Machine Learning Engineer Nanodegree

Capstone Proposal

Amitabha Chakravarty
4 July 2017

Kaggle Competition - **Planet: Understanding the Amazon from Space**

Domain Background

Our planet is loosing valuable resources every minute. Every minute an area of forest the size of 48 football field is disappearing from Amazon basin due to deforestation. Deforestation has caused a large scale of devastating effects – reduction of biodiversity, habitat loss, climate change to name a few. To counter the effect of loss of such forest area we need data that will pin point areas of human encroachment. This will help local governments and local stakeholders to respond quickly and effectively.

For this purpose, daily imagery from Planet, designer and builder of world's largest constellation of Earth-imaging satellites, will be used along with its Brazilian partner SCCON. They want Kagglers to label the satellite image chips with atmospheric conditions and various classes of land cover and land use. Resulting algorithms will help the global communities understand better where and how deforestation is happening and ultimately how to deal with it.

Problem Statement

Image (chip) data is collected from Planet's full-frame analytic scene products using 4 band satellites in sun-synchronous orbit. The data is labeled using Crowd Flower platform and a mixture of crowd-sourced labor. There are class labels like 'Cloudy', 'Partly cloudy + Primary', 'Shifting cultivation + primary' and so on. There are a significant number of cloudy scenes with complete to partial cloud coverage and also hazy conditions. There are also clearly shown images of 'Primary Rain Forest', 'Water (River and Lakes)' and 'Habitation' etc. The job is to come up with a data analysis model and use these labeled data for training purpose. There will be a set of images with no label and the model has to be tested against the unlabeled data.

Datasets and Inputs

Training – a list of training images and their corresponding labels are presented. Each image is a 256x256 pixels corresponding to a real area of 221.7 Acres.
Testing – a list of images is given without labels and the solution will submit the predicted labels against them.
We have around 40000 images for training and 40000 images for testing. There are both

jpg and 4-band tiff images for training and testing. The majority of the data set is labeled as "primary", which is shorthand for primary rainforest, or what is known colloquially as virgin forest. There are also areas representing 'Water', 'Agriculture', 'Road', 'Cultivation' and so on. The labels are presented as applicable to the ground reality with multiple keywords like – 'Agriculture/pasture + primary + partly cloudy'.
There are two kinds of images - a "hard" and an "easy" set. The easy set contained scenes that are easier-to-identify labels like primary rainforest, agriculture, habitation, roads, water, and cloud conditions. The harder set of data was derived from scenes to represent shifting cultivation, slash and burn agriculture, blow down, mining, and other phenomenon. A disclaimer is mentioned on the competition page that some training labels could be wrong as there could be labeling error. Overcoming the inaccuracy would be a challenge for this particular completion. There is also significant cloud covering for some of the images. Detecting scenario on ground in presence of cloud would be particularly challenging.

Solution Statement

The goal of this project would be to develop and algorithm to detect the type of scene represented by the testing images with varying atmospheric conditions and various classes of land cover/land use. There are several approaches to analyzing an image and classify the scene based on prior knowledge. One such approach would be finding PCA (Principal Component Analysis) components and classify using SVM (Support Vector Machine). However, with recent success of deep learning, Convolution Neural Network (CNN) is another very effective technique to analyze image and classify its content based on prior learning. We will go with CNN approach here for its high success rate.
CNN would be used to train and detect the correct label for the ground scenario represented by the amazon image chip. CNN outperforms many image recognition algorithms in ImageNet dataset, The CNN algorithm consists of many convolution operations followed by pooling operation to feed finally to fully connected layer. Finally, the prediction comes out of the classification layer connected next to the fully connected layer.

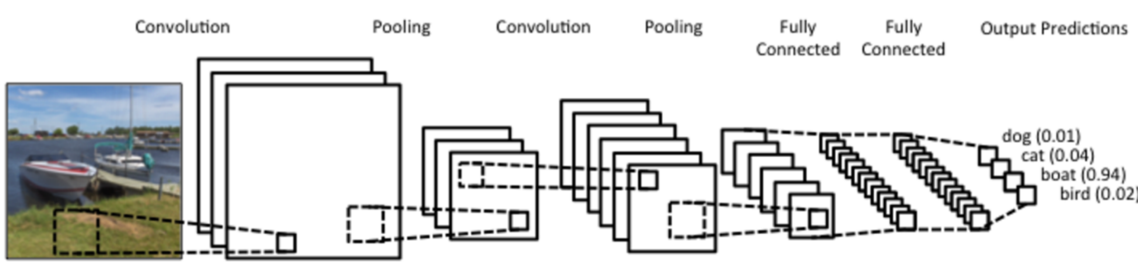

Image courtesy - http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Benchmark Model

Each Kaggle competition has a leaderboard where competitors' scores are arranged in descending order. We can look at the range of results obtained by the leading competitors and arrive at a benchmark score for the algorithm.

After looking at the leaderboard for current competition at https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/leaderboard we see that about 700 teams are competing and here are some of the score ranges –

~400 of them scored 90% or more
~350 scored 91% or more
~215 scored 92% or more
~56 scored 93% or more – TOP Tier with highest score being 93.385%

We can choose a benchmark score of 91% as half of competitors have crossed that mark. We need to better that score to consider our model as successful.

Evaluation Metrics

Two kind of metrics will be utilized. One is the 'Accuracy' of predicted label against validation data. The other is the 'F1 Score' which is defined as follows –

F1 = 2 * (precision * recall) / (precision + recall)

This score is based on the 'true positive' and 'false positive' as calculated as part of validation.

Project Design

Our Convolutional Neural Network will process images through multiple layers and finally produce a classification label. However, it is not a single class. From the training data it is seen that multiple labels are applied against each image. Here are some samples –

```
image_name                                       tags
0    train_0                              haze primary
1    train_1            agriculture clear primary water
2    train_2                             clear primary
3    train_3                             clear primary
4    train_4  agriculture clear habitation primary road
```

So, we have to predict more than one class instead of just one. We can adopt something similar to one-hot-encoding for the labels. We can convert the tags to a fixed length sparse vector where only the corresponding places of labels that are found we have to insert 0 , otherwise 1.

For example – we can have following vector for train_0 image –

```
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  1.  0.  0.  0.  0.]
```

where 1 in 11th place and 13th place indicate presence of labels 'haze' and 'primary'.

With output layer we can apply softmax function to convert the actual output to probability. Since it is a multiple label prediction we need to set a threshold to detect strong presence of a label.

The image sizes are 256x256. Since our computation capability is limited, we would resize the images to either 64x64 or 32x32 and so on based on achieved success rate. We will start with normalized jpg images only and based on success will consider tiff images if required.

Different optimizers like ‘AdamOptimizer’ or ‘GradientDescent’ will be experimented. We will optimize the cross-entropy of the predicted vs actual label-vector. Data will be analyzed in batches and best models need to be check-pointed. We will use a part of the training data to validate the correctness of the model being developed. Standard techniques of ‘split and validation’ would be utilized to separate training set from validation set.

We will be experimenting on various parameters like ‘learning-rate’, ‘pooling frequency’, convolution strides and k-value, number of layers etc.

Finally, the testing dataset would be used to predict final score for the derived model.