Machine Learning Engineer Nanodegree

Capstone Proposal

Amitabha Chakravarty
1 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 partnet 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.

Benchmark Model

As there is no benchmark model presented by the competition a part of the training data itself would be used to validate the correctness of the model being developed. Standard techniques of ‘split and validation’ would be utilized. The labelled training dataset would be divided into a training and a validation set according to a predetermined ratio (say 8:2). The trained model would be validated against the validation data and the model would be improved iteratively.

Evaluation Metrics

Two kind of metrics will be utilized. One is the ‘Accuracy’ of predicted label against validation data. At each step of training MSE (means square error) would be calculated corresponding to predicted and actual label and the model would be improved.

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

Convolutional Neural Network (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.

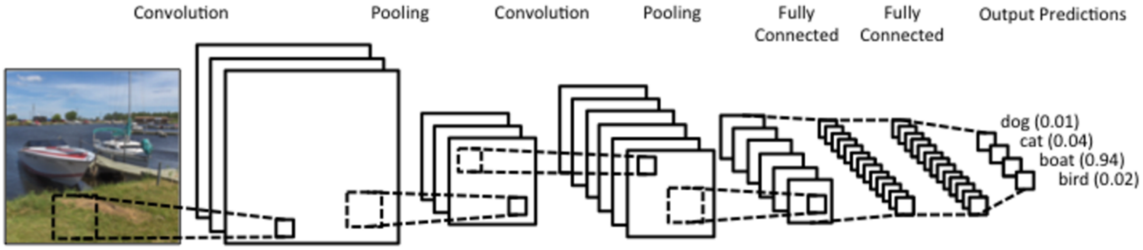

Image courtesy - http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/