ML Engineer Capstone Proposal

Using Convolutional Neural Networks to Identify Cactus Species

Adriano Torres
adrianortorres@gmail.com

Udacity
Nanodegree Machine Learning Engineer

1 Domain Background

Climate change is a growing concern in today's political and economical landscape. As forests, and vegetation in general, play a considerable role in the process of carbon exchange, it is important that we devise technology to preserve them. Large ecosystems, such as the Amazon forest or the African Savanna, are under a growing threat due to anthropic effects on climate change. Not only is human activity indirectly impacting these biomes, but active deforestation is prone to take place in these forests, for it can be difficult to monitor these vast regions.

In this context, machine learning can be a powerful tool in the development of autonomous systems whose goal is to monitor landscape changes over time, as well as providing a means for identifying persons or groups who engage in illegal deforestation activites..

In this work we aim to apply elementary Convolutional Neural Network methods for computer vision, taregeting to identify a particular species of cactus - the Neobuxbaumia tetetzo - from aerial images. The goal of this project is to illustrate how to build a fundamental component of an eventually applicable system in autonomous surveillance. The work developed in this example could be extended to recognize multiple species, and it could also be augmented by more powerful techniques to monitor changes throughout time.

The topic is personally motivating to the author, as he is from Brazil, where most of the Amazon jungle, the largest tropical rainforest on Earth, is located, and because illegal deforestation is a prevailing issue since the early 1900s.

2 The problem

Our problem consists in, given a dataset of aerial images, determining whether there is a cactus species in the image. Provided that our dataset is correctly labelled, then our problem is, by definition, quantifiable. It is also possible to measure our model's accuracy by comparing its predictions with the actual labels attached to each picture. A metric for measuring the model's performance is stated in Section 6. Finally, our method is replicable, as any one in possession of the data and the code can run the predictions themselves. Furthermore, once the network is trained and performing according to our standards, it can theoretically be run to predict on new input data, provided such data is in the same format as the one used to create and train the model.

3 Dataset

The dataset we will be using is a subset of a more comprehensive dataset created by researchers responsible for creating the VIGIA project in Mexico. The dataset has been simplified by Kaggle. Our working dataset has been resized to 32 x 32 images. The actual data is composed of two columns:

1. id: A string composed of lowercase letters and digits , which uniquely identifies each picture. Each string is followed by the jpg extension. Thus, a sample id would be 011adabb3831de1deaf542b82ac870d1.jpg

2. has_cactus: A label which determines whether a cactus is present in the picture. In the positive case, the value is 1, and 0 otherwise.

The data has been previously split between testing and training samples. The training folder contains 17500 pictures, whereas the testing set has 4000 pictures.

4 Solution outline

Our solution will be based on Convolutional Neural Networks, or CNN's, implemented using the Keras framework for Python. The rationale for these options is based on the facts that, in comparison to other famous neural network libraries in python (i.e. TensorFlow and PyTorch), Keras has a smoother learning curve. The reason behind the choice for CNN's is less obvious to the writer at the moment, but most queries in Google Scholar and Kaggle on image recognition and classification return CNN models with high relevance rates. Hopefully the reason will be clearer as we develop and train the model, as well as in further courses and study on the subject.

5 Benchmark Models

At the time of this writing, five out of the 132 solutions submitted have reached maximum score. Since it is thus possible to attain maximum score of 1.0000, our model will be benchmarked against the highest score possible. Our goal is to attain 0.8000 or better.

6 Evaluation Metrics

As stated in the Evaluation section of the Kaggle Competition, projects are going to be evaluated based on Receiver Operating Characteristics - ROC. The method basically consists in calculating the recall, or True Positive Rate, versus the complement of the specificity, which is also referred to as fall-out. This effectively produces a function whose input is the fall-out, and the output is the sensitivity. Based on this, our model is going to be assigned a score which corresponds to the area under the ROC curve.

7 Project design

The general workflow of our project will consist in an initial section where we will understand and visualise the dataset. Since there are only two columns in our dataset, and we are determined to use CNN's, we will begin by studying CNN models, reading examples, and understanding the underlying theory behind them. After we have developed a decent conceptual understanding, we will proceed to implement a basic one using the Keras library. At this point, when we have a working prototype, we will iterate, tuning hyperparameters, such number of layers, activation methods, and number of epochs. After we have reached the desired score of 0.8000, we will compare our solution to the top scorer's, and try to gain insight on how to make the simplest possible tweaks in order to enhance our performance.