Machine Learning Engineer Nanodegree

Capstone Proposal

Luke(Xu) Lin
January 15th, 2018

Dogs vs. Cats Redux: Kernels Edition

Domain Background

There are lots of things that is easy for humans, dogs, cats, but your computer will find it a bit more difficult. Web services are often protected with a challenge that's supposed to be easy for people to solve, but difficult for computers. Such a challenge is often called a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) or HIP (Human Interactive Proof). HIPs are used for many purposes, such as to reduce email and blog spam and prevent brute-force attacks on web site passwords.

Identify photographs of cats and dogs. This task is difficult for computers, but studies have shown that people can accomplish it quickly and accurately. Kaggle has redux version for this. 1314 teams have been participated in this project and the leaderboard can be viewed here. Also there are some academic research about this project, like here

My personal motivation for this is improving my web crawler perfomance. Lots of time my web crawler will stop when it comes with CAPTCHA. I beleive this project will help me have betting understanding how to deal with this situation.

Problem Statement

Our basic task is to create an algorithm to classify whether an image contains a dog or a cat. The input for this task are images of dogs or cats from training dataset, while the output is the classification accuracy on test dataset. TFLearn might be applied as one of the solution method.

We will have a data set (images) with label indicated whether it is a dog or cat, we expect we can come up with an algorithm to calculate the probability that the image is a dog (1 = dog, 0 = cat).

We will use Log Loss to value the result.

Datasets and Inputs

The data is coming from here The train folder contains 25,000 images of dogs and cats. Each image in this folder has the label as part of the filename. The test folder contains 12,500 images, named according to a numeric id. For each image in the test set, you should predict a probability that the image is a dog (1 = dog, 0 = cat).

Solution Statement

To tackle the problem described in Problem Statement, we will use TFLearn to build a few neural network, and feed the train data into it, get log loss as result. Then compare those neural networks by log loss result, time consumed.

Benchmark Model

Accroding to Kaggle Leaderboard, the first position of 1314 teams was taken by team "Cocostarcu", their Log Loss score is 0.03302. I hope I can be the top 20% of this project, which mean the Log Loss result should be lower than 0.08167.

Evaluation Metrics

The project is evaluated by log loss

Project Design

- Download TFLearn and project data set
- try to design a few neural network
- feed train data to neural network
- get log loss result, time consumed
- compare performance of those neural networks by log loss result and time consumed

Tools and Libraries used: Python, Jupyter Notebook, pandas, scikit learn, TensorFlow, Keras. Other libraries will be aded if necessary.