CAPSTONE PROJECT RAJ DAYAL NATARAJAN

MACHINE LEARNING NANODEGREE DATE: OCT 1ST, 2017

DOMAIN BACKGROUND:

The problem I chose to tackle for my project is a Kaggle competition which is currently underway. The project name is “Cdiscount’s Image Classification Challenge”. The following is the background. Cdiscount.com is France’s largest non-food e-commerce company with a revenue of 3 Billion Euros. The company sells everything from TV’s to trampolines and has about 30 million products which are up for sale by the end of the year. The company now faces a challenging problem of ensuring that all these product are correctly classified so that the website can be streamlined to provide a good user experience to the customers.

PROBLEM STATEMENT:

The problem seen by Cdiscount as mentioned in the background is as follows. With an ever growing array of products being sold by Cdiscount, the company faces the challenge of correctly classifying these products into distinct categories so as to improve the customer’s online experience with Cdiscount.com.

Currently, Cdiscount is using Machine Learning techniques to infer the classification of the product from the description of the product. However, the company thinks that the next major quantitative improvement in classification accuracy can only come from using images to classify the products in addition to the description.

DATASETS AND INPUTS:

The data set has been provided from its internal catalogs and has about 9 million products with 15 million images with a standard resolution of 180x180. There are approximately 5000 output categories. In short, it would be a big data, multi-class classification problem.

The data set has been provided in BSON format - binary encoded serialization of JSON like documents.

Details: (Note: From the competition description)

Training Data set:

“ *(Size: 58.2 GB) Contains a list of 7,069,896 dictionaries, one per product. Each dictionary contains a product id (key: \_id), the category id of the product (key: category\_id), and between 1-4 images, stored in a list (key: imgs). Each image list contains a single dictionary per image, which uses the format: {'picture': b'...binary string...'}. The binary string corresponds to a binary representation of the image in JPEG format.”*

Testing Data Set:

*“Size: 14.5 GB) Contains a list of 1,768,182 products in the same format as train.bson, except there is no category\_id included. The objective of the competition is to predict the correct category\_id from the picture(s) of each product id (\_id). The category\_ids that are present in Private Test split are also all present in the Public Test split.”*

*Category names:*

*“Shows the hierarchy of product classification. Each category\_id has a corresponding level1, level2, and level3 name, in French. The category\_id corresponds to the category tree down to its lowest level. This hierarchical data may be useful, but it is not necessary for building models and making predictions. All the absolutely necessary information is found in train.bson.”*

SOLUTION STATEMENT

BENCHMARK MODEL

EVALUATION METRICS

PROJECT DESIGN