Capstone Proposal Andrea Balzano

Machine Learning Engineer Nanodegree 6th April 2015

Project overview

The e-commerce economy is fast grooving and the expectation from the customer is higher than ever before.

In this project, I dive deep in the fashion industry where the market is crowded with online shops and the offering is very similar. Nowadays price is not driving preference as much as customer service and presentation, companies can benefit from machine learning algorithm that can identify clothing attributes independently, this has been mostly used to achieve precision in search engines.

In this project, I want to analyse the attribute of the image before the ones of the clothing item portrayed, to understand customer selection and address problems in sales.

Problem Statement

The goal is to create an agent able to identify customer preferences for particular style, accessory or colour as well as identify the attributes of best/worst selling items. The task involved are the following:

1. Download and preprocess the images.

2. Train a classifier that can extract attributes from images.

3. Apply clustering for customer suggestion.

4. Apply clustering for best / worst selling item.

5. Train a classifier to predict if an item will be a best seller.

The final application is expected to be used by the marketing team during campaign creation.

I expect to show these agents can provide useful information to save on campaign costs, especially in shooting and image editing, by suggesting poses and on product selection by predicting customer feedbacks on the item advertised.

Datasets and Inputs

The dataset I will use is a subset of ‘deepfashion’, information about the author can be found here: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion/InShopRetrieval.html

The dataset is public for non-commercial purposes and contains different clothing items and multiple poses per item.

This can be used in conjunction with customer information extracted from weblogs such as if the image is selected, added to cart, purchased. I will prepare a file simulating this information.

Metrics

The F1 score will be the metric used for the classifiers with the use of a confusion matrix to identify the general performance of the agents.

Mahalanobis distance will be the metric used for the clusters.

Project Design

I will approach preprocessing by applying PCA and ICA to the images, I expect that PCA will be a good preprocessing step for the CNN that will classify the images.

I am curious to see if ICA is able to identify the attributes of the clothing items and if I can use that information combined with customer information to apply clustering and therefore extract customer preferences.

I will try to combine the output of the CNN with customers into a regression to predict best / worst selling.