April 3rd, 2022

CSCI E-25, Computer Vision

Project Proposal

Automatic category identification and filter mapping for Products (Focus: Fashion Ecommerce Categories)

By: Sakhuja, Sahil

Introduction

In e-commerce portals, we have thousands, if not millions, of products. With a view to tackle reducing attention spans of customers and lower exclusivity in product portfolios, the capability to have improved searchability within the catalogue is a critical desired outcome for companies operating in this space.

There are primarily two ways that consumers search for products in any e-commerce portal, namely:

1. Using the Search option which performs a free-text search in the names and descriptions of the products
2. Traversing through the categories and sub-categories on the website and then using provided Filters to restrict the number of displayed items

In this project, I would like to focus on improving results for the 2nd route i.e. creating an automated model to help e-commerce site owners & administrators to maximize accuracy of categorization and filtering.

Categorization and filtering of products is based on inputs made into a cataloguing system. This creates 2 points of failures:

1. Incorrect or insufficient category mapping primarily due to:
    a. Manual blunders: Men's trousers categories under Women's trousers
    b. Insufficient breadth: A pair of unisex shoes only categorized under Men's shoes
    c. Insufficient depth: A pair of Sneakers only categorized into the top category as shoes and not sub-categorized as Sneakers
2. Insufficient data input for filters eg. Collar type not updated on a Shirt

These issues would lead to inefficiencies in product search by customers.

Project Goal

I would like to focus my project on developing a model which can, given an image, be able to recommend the right category / sub-category for a product as well as recommend the relevant filter options to be enabled for the product and the value thereof. To limit the scope, I am focusing the project only on the Fashion category.

Hence, I will be training ML models for 2 distinct outcomes:

1. Recommendation of a category for the product in the image.
2. Recommend different filters, and the respective values thereof, for the product in the image.

Dataset

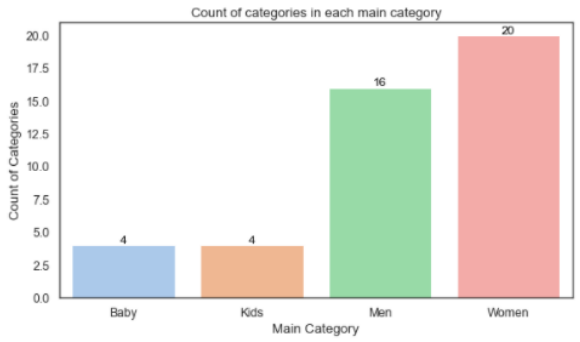
I have used web-scraping to fetch the data from a prominent Fashion e-commerce portal in Germany and have extracted the following artefacts:

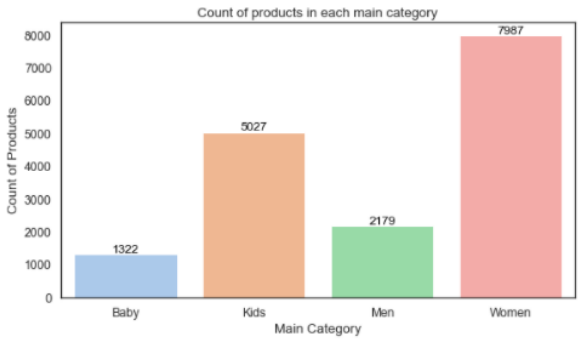
1. Images – only "still-life" images i.e. images which are of the product itself and do not include any models.
2. Category / Sub-Category – The category and sub-category that those products are currently tagged in. These are the "true labels".
3. Filters and Filter Values – the attributes of the product which are available as filters on the images and the values thereof. These are the "true labels".

The data is stored as follows:

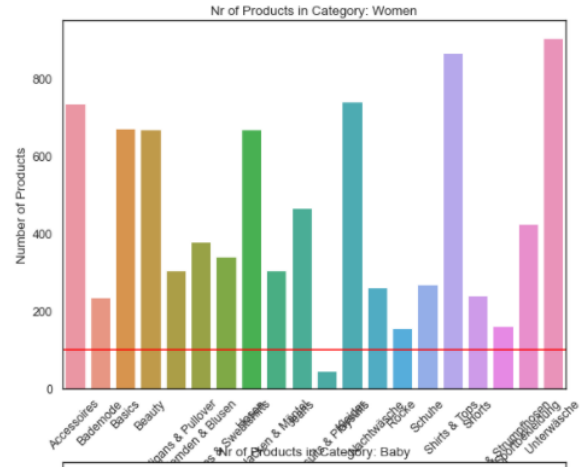
1. 4 files with relevant data and mappings, namely:
    a. Categories: Dataframe with information on category mappings i.e. main categories (Men, Women, Baby and Kids) and sub-categories within these (eg. Shirts, Trousers, Dresses, etc.).
    b. Items: Dataframe with information on each item listed on the e-commerce portal (snapshot as on 12th / 13th / 14th March, 2022) with details including: Item code, Category Id, Image URL, Image Extension, etc.
    c. Filters: Dataframe containing information on filters available in each sub-category eg. Colour, Pattern, Sleeve Length, etc.
    d. Filter Values: Dataframe containing the Item code, the fitler Id and the value of the Filter for that item.
2. Images of each item – named in the format {Portal_Name}_{Item_Code}.{Extension}.

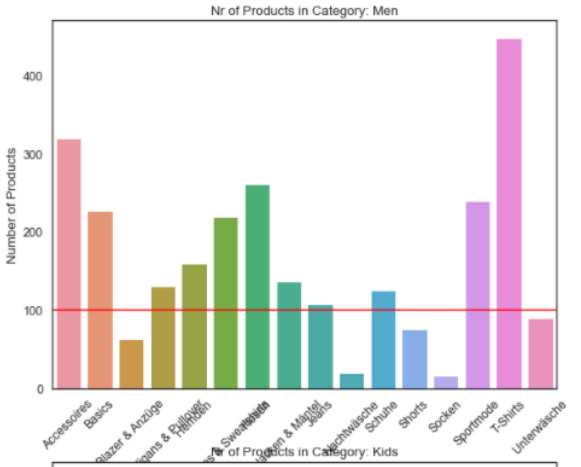
The data consists of 16,473 products (with images) spread across 4 main categories – Baby, Kids, Men and Women.

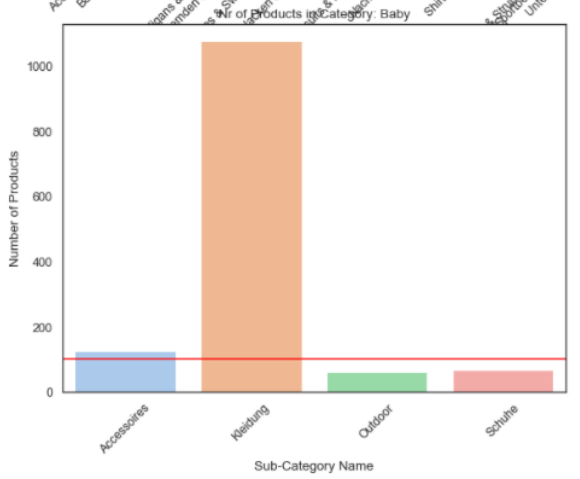


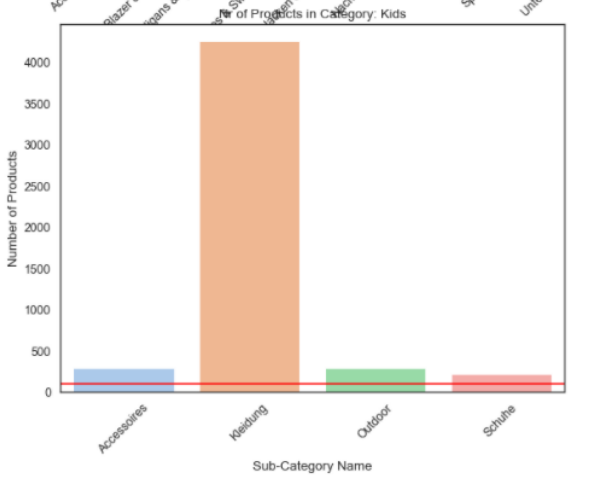


Within each category, there are also sub-categories present,



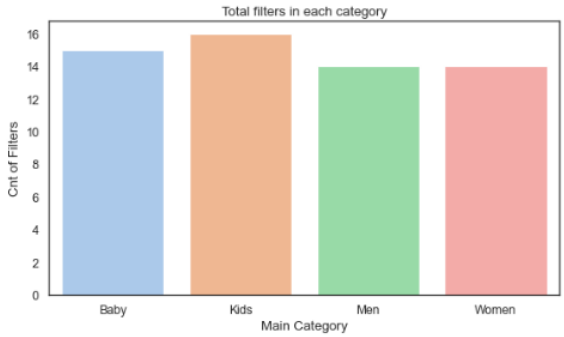


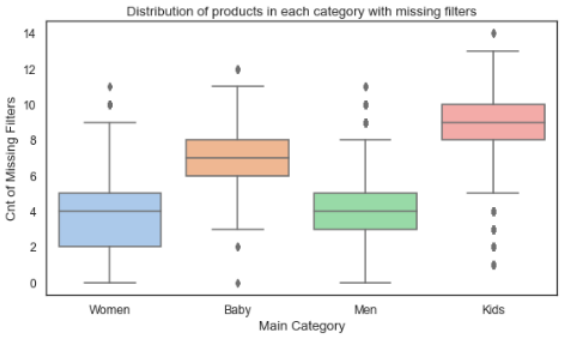




I have chosen to only consider sub-categories which have a minimum of 100 products – hence, this translates into a total of 36 categories which can be used as outcome labels. The remaining categories and products within them would not be used for the purpose of the project.

Additionally, there are also various filters available in each category. On an average, each main category has about 15 filters. These filter options seem to be Missing at Random on some of the observations from the data set and could be the result of either non-applicability (i.e. the filter option is not relevant for the product) or erroneous data entry. In the scope of the project, we would only be using the observations which have the filters mapped. The remaining values would be excluded.





Pre-Processing

All images are of the size 453X302 with 3 colour channels. As part of the web scraping, I have been able to get images which are "still life" i.e. there are no models in the images and hence, they are directly usable for the purpose of the project.

Some example images:

ML Modeling

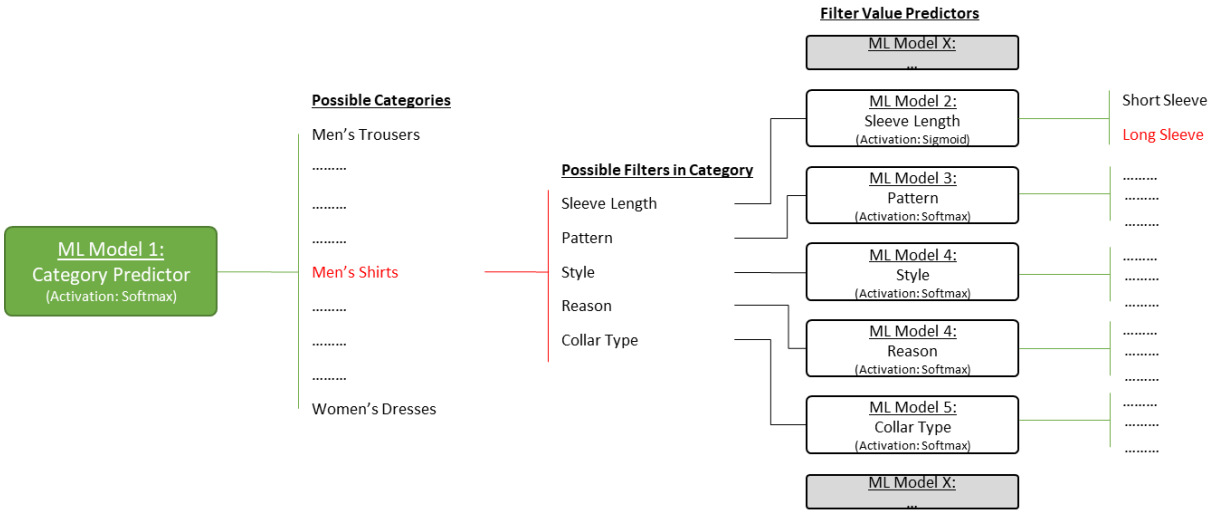
I will be aspiring to train multiple models for different outcomes. I would be running these ML models in a pipeline to which will follow the below steps for every input image:

1. **Step 1: ML Model for Category** – A single multi-classification model to predict a category Egs. Men's Shirts

2. **Step 2: Choose Filter Options -** Depending on the category, I would be able to determine which filter options are possible for the product – For eg. Collar Style would be relevant for Men's Shirts but not for Men's Trousers

3. **Step 3: ML Model(s) for Filter Values –** Each filter option identified in previous step would have it's own model for prediction of the value for that filter, Egs. Collar Style = Button-Down; Sleeve Length = Half-Sleeve, etc. These models would be either binary classification or multi-class classification depending on the number of possible values for that filter. Additionally, we will be using an "abstain" model in this case i.e. no prediction would be made if the prediction probability is below a certain threshold – this hyperparameter will be tuned over the range of possible values between (0, 1).

Hence, I will need to train multiple models and use them in a pipeline. An example of how this would work is given below.



I will try to train these models as follows:

1. Convolution Layer using Transfer Learning – I will experiment with some pre-trained models and try to use them for transfer learning to get higher accuracy in the convolution layers. This will also help in reducing the need for data augmentation.
2. Classification models using fully connected layers – On the output of the convolution layers, I will train the different classification models using a dense fully connected neural network.

Depending on the complexity and accuracy I am able to achieve, objective (1) i.e. category prediction would be the most important target outcome for the purpose of the project and objective (2) i.e. filter options prediction would be a secondary / optional outcome.