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

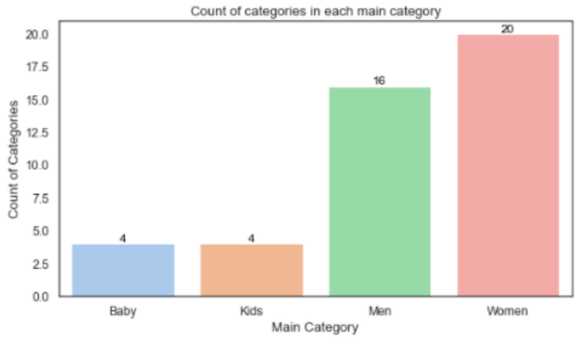
I have used web-scraping to fetch the data from a prominent Fashion e-commerce portal in Germany and have extracted the following artefacts:

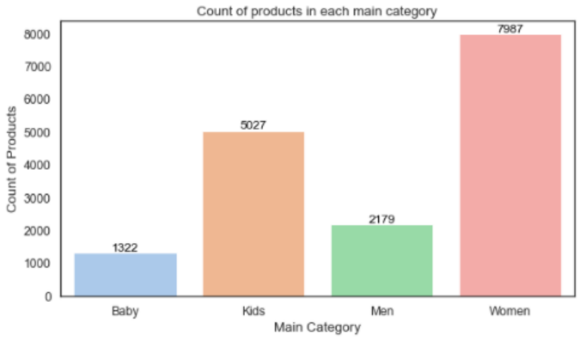
1. Images – only "still-life" images i.e. images which are of the product itself and do not include any models.
2. Category / Sub-Category – The category and sub-category that those products are currently tagged in. These would be considered as the "true labels" for the purpose of the project.
3. Filters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

The data is stored as follows:

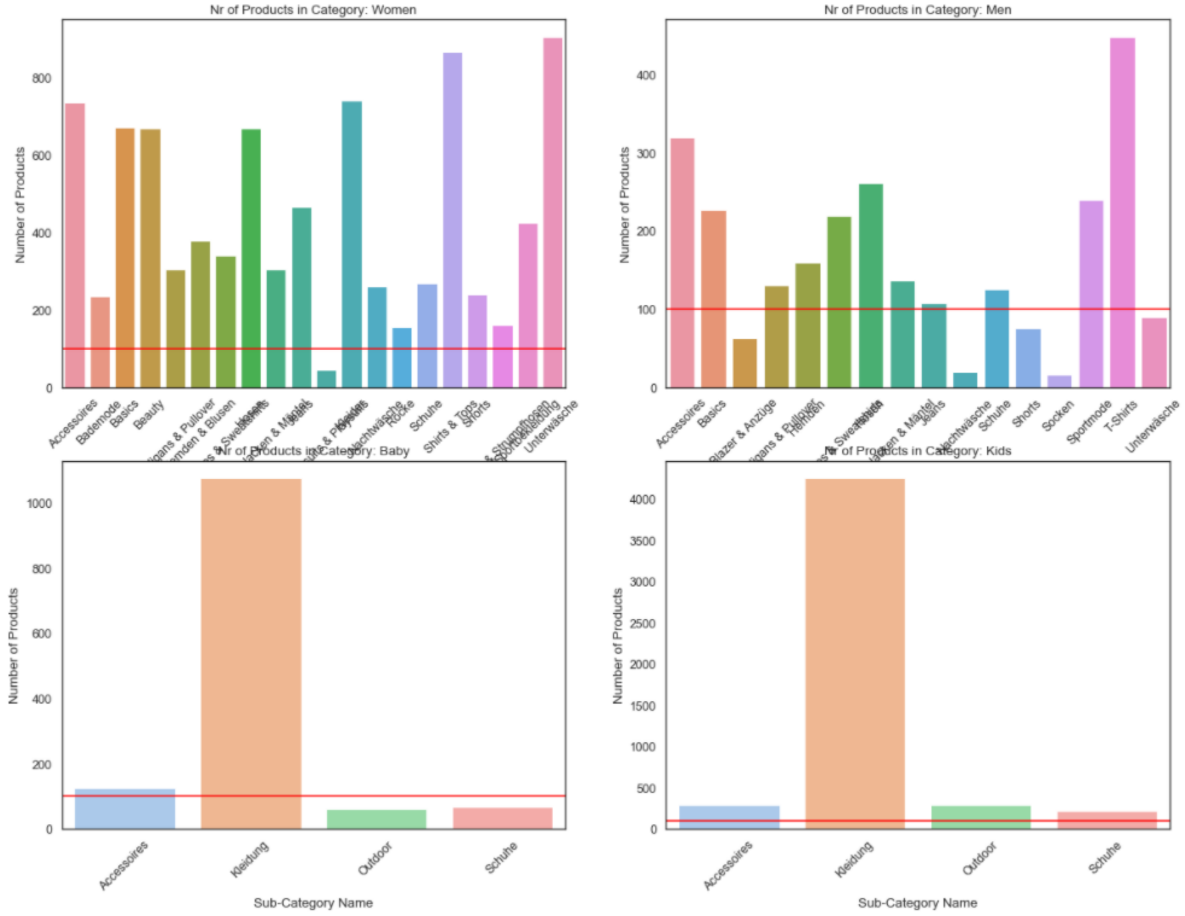
1. 4 files with relevant data and mappings, namely:
    a. Categories: Dataframe with information on category mappings i.e. main categories (Men, Women, Baby and Kids) and sub-categories within these (eg. Shirts, Trousers, Dresses, etc.).
    b. Items: Dataframe with information on each item listed on 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.



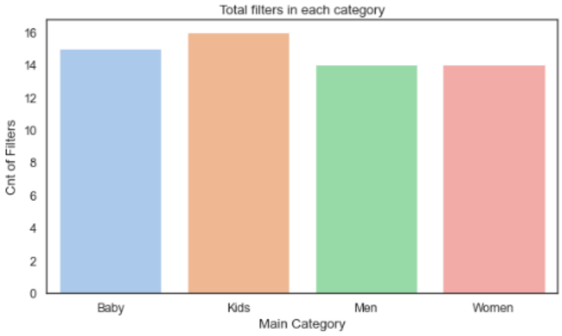


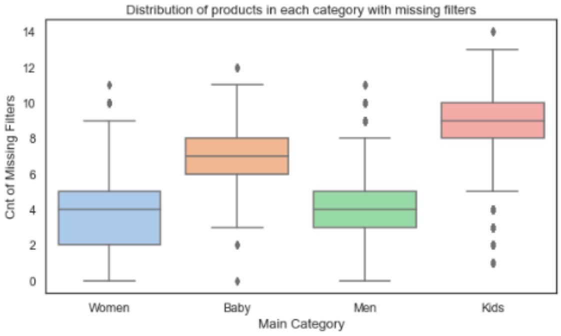
Within each category, there are also sub-categories present,



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.

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 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.





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.

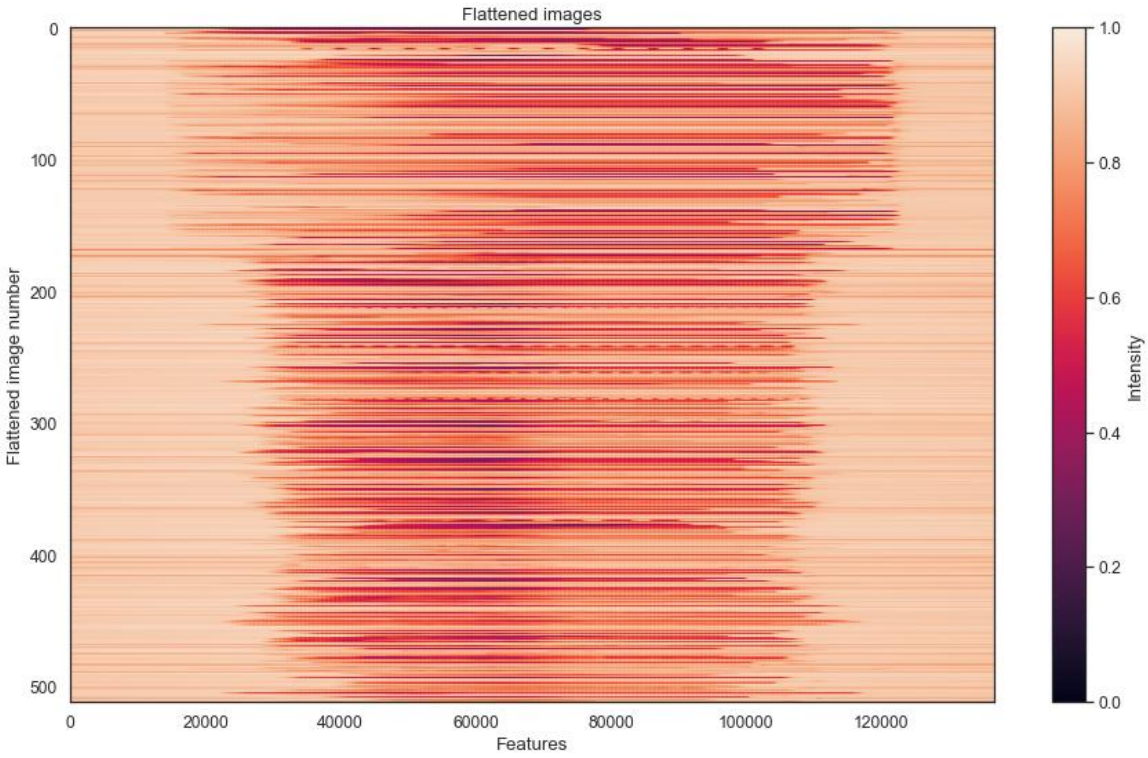
Some example images:

Feature Engineering & Extraction

Since the images are quite large, when converting them to 1-d arrays, the resulting feature space includes 410,418 features. Considering that the total training images are only around 16,000, I will be exploring techniques for reducing the dimensionality of the feature space in order to avoid over-fitting. Some of the dimensionality reduction and regularization techniques that I would be using in order to pre-process images are as follows:

1. Converting to grayscale – Since the project is more focused on object identification in the images, the colour channel information would be redundant to the outcome. Hence, I will convert all images to grayscale before processing – this will reduce the features to 136,806.

2. Removing background padding – All images have grey backgrounds, especially around the borders, which do not hold any relevant information (as showcased in the plot of a sample of the features below) for the purpose of my project.



On using Variance threshold, it is possible to reduce this feature space to approx. 51,000 features – this number may change at a later point since I have only performed this on a sample for now.

3. Data Augmentation – I would also be using data augmentation techniques to increase the number of images available for training and testing.

4. Edge & Corner detection – Since the project outcome relies a lot on the overall shape of the objects in the image, I will be experimenting with different edge / corner detection algorithms and using them as features in the modelling process.

ML Modeling

I will be aspiring to train multiple models for different outcomes, namely:

1. **Category prediction** – I will train a multi-class classification convolutional neural network for recommending the category for the product from the 36 possible outcome categories.

2. **Filter options prediction** – For the second part of the project focused on recommendations of filter options, I will first need to identify the relevant filter options for a product. On initial observations from the data, I can observe that there are a lot of products with missing filters. Hence, it may be assumed that not all filters are applicable for each product in the category. So, I will first be making a prediction, via multiple binary classification models, to establish if a specific filter applies to a product or not. And then, for each filter that has been predicted as being viable for a product, I will use a multi-class classification model to predict the option value.

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.