Imaterialist challenge - Report

Mick van Hulst Dennis Verheijden Roel van der burg Brian Westerweel
Joost Besseling

May 17, 2018

1 Exploratory Data Analysis

The dataset that the challenge provides consists of a training, test and validation set, which in turn consist of:

- 1.014.544 training images;
- 228 labels in the training dataset;
- 39.706 test images;
- 9.897 validation images;
- 225 labels in validation set.

We're interested in the distributions of the data such that we can prepare our models accordingly. Figure 1 and 2 visualise the distribution of the labels for the training and validation dataset respectively. We observe that the labels aren't evenly distributed, but the distribution between the training and validation dataset seems comparable.

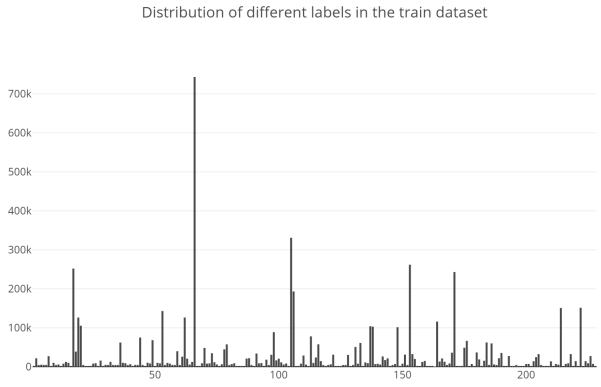

Figure 1: Data distribution training set

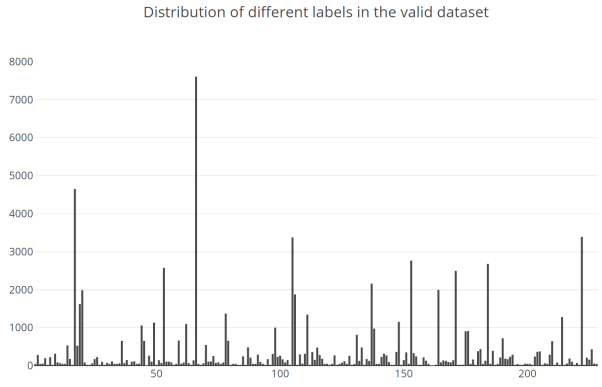

Figure 2: Data distribution validation set

To get a feeling for the dataset we also visualise some images and observe that for some labels there are no clear characteristics that define the corresponding label (see Figure 3). From this observation we conclude that it would be hard for a human to differentiate between the labels without knowing what the labels mean exactly.

We believe that there's some structure to the labels, meaning that there might be some tree-like structure that explains the different combinations between the labels, however, this is not something that the challenge provides so we cannot use this to e.g. define weights when classifying. This is something we'll have to define ourselves.

Figure 3: Example image label 24

Source: `https://www.kaggle.com/shivamb/imaterialist-fashion-eda-object-detection-colors`