

Semantic Segmentation of Fashion Images Using Feature Pyramid Networks

John Martinsson

RISE Research institutes of Sweden

john.martinsson@ri.se

Olof Mogren

RISE Research institutes of Sweden

olof.mogren@ri.se

Abstract

In this work, we approach the problem of semantically segmenting fashion images into different categories of clothing. This problem poses particular challenges because of the importance of both textural information and cues from shapes and context. To this end, we propose a fully convolutional neural network based on feature pyramid networks (FPN), together with a backbone consisting of the ResNeXt architecture. Our experimental evaluation shows that the proposed model achieves state-of-the-art results on two standard fashion benchmark datasets, and a qualitative study verifies its effectiveness when applied to typical fashion images. The approach has a modest memory footprint and can be used without a conditional random field (CRF) without much degradation of quality which makes our model preferable from a computational perspective. When comparing all methods without a CRF, our approach outperforms all state-of-the-art models on both datasets by a clear margin in all evaluated metrics. In fact, our approach achieves a higher accuracy without the CRF than the state-of-the-art models using CRFs.

1. Introduction

Analysing trends is an important strategic activity in the fashion industry, which is increasingly becoming a task of identifying trend setting individuals, following them on blogs and social media platforms. Fashion is to a large extent communicated with images, and going through large quantities of photos is one of the key tasks. Visual data is useful for successful fashion forecasting. Al-Halah, et.al. [1] used image features produced by a convolutional neural network (CNN) model trained for image classification to perform fashion style forecasts. We assume that the richer information given by semantic segmentation is beneficial of such downstream tasks. Having the right tools at hand to aid this work can allow analysts to work more effectively, and extracting semantically rich representations from images can be such a tool, providing detailed information to sort through the massive stream of data.

Figure 1: From left to right: the input image, the ground truth segmentation, the predicted segmentation (ResNeXt-FPN), the prediction with CRF, and the incorrectly classified pixels (shown in black) for two top scoring test data image predictions from refined Fashionista.

We consider an important part of such a toolchain: semantic segmentation of fashion images. That is, the division of the image into different regions of clothing.

Figure 1 shows an example image, with the ground truth segmentation, and predicted segmentation maps produced by our approach. Each pixel in the image is classified, and the output is a semantic analysis showing which clothing items are present, where in the image these are, and what shape they have. Shapes of clothing items are an important feature for fashion analysis. For example, a hat could have many different shapes, and the shape of a hat may go in and out of fashion over time.

We use a feature pyramid network (FPN) [8] with a ResNeXt [11] backbone for the semantic segmentation of fashion images. The feature pyramid structure makes the model more robust against images of different scales, and allows both high and low level features to be used in the prediction of the semantic segmentation map. The filters learned by the early layers in a CNN usually resemble Gabor filters or color blobs [13] and such low level features have been shown to improve the accuracy of cloth-

