Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction

Charles Corbière[1], Hedi Ben-Younes[1,2], Alexandre Ramé[1], and Charles Ollion[1]

[1]Heuritech, Paris, France
[2]UPMC-LIP6, Paris, France
{corbiere,hbenyounes,rame,ollion}@heuritech.com

Abstract

In this paper, we present a method to learn a visual representation adapted for e-commerce products. Based on weakly supervised learning, our model learns from noisy datasets crawled on e-commerce website catalogs and does not require any manual labeling. We show that our representation can be used for downward classification tasks over clothing categories with different levels of granularity. We also demonstrate that the learnt representation is suitable for image retrieval. We achieve nearly state-of-art results on the DeepFashion In-Shop Clothes Retrieval and Categories Attributes Prediction [12] tasks, without *using the provided training set.*

1. Introduction

While online shopping has been an exponentially growing market for the last two decades, finding exactly what you want from online shops is still not a solved problem. Traditional fashion search engines allow consumers to look for products based on well chosen keywords. Such engines match those textual queries with meta-data of products, such as a title, a description or a set of tags. In online luxury fashion for instance, they still play an important role to address this customer pain point: 46% of customers use a search engine to find a specific product; 31% use it to find the brand they're looking for [1]. However, those meta-data informations may be incomplete, or use a biased vocabulary. For instance, a description may denote as "marinière" a long sleeves shirt with blue/white stripes. It then appears crucial for online retailers to have a rich labeled catalog to ensure good search. Moreover, these search engines don't incorporate the visual information of the image associated to the product.

[1]http://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-opportunity-in-online-luxury-fashion

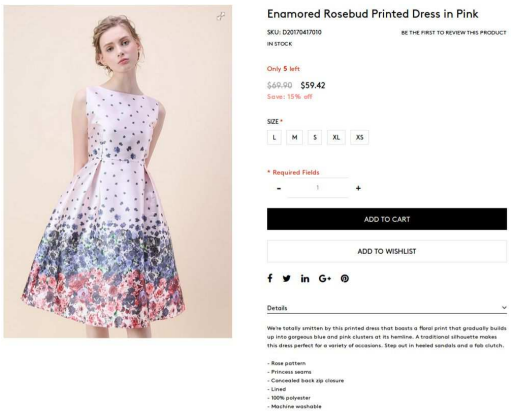

Figure 1. Our dataset is composed of images and a few associated text descriptors such as their title and their description

Computer vision for fashion e-commerce images has drawn an increasing interest in the last decade. It has been used for similarity search [20, 11, 21, 18], automatic image tagging [10, 2], fine-grained classification [5, 12] or N-shot learning [1]. In all of these tasks, a model's performance is highly dependent on a visual feature extractor. Using a Convolutional Neural Network (CNN) trained on ImageNet [4] provides a good baseline. However, there are two main problems with this representation. First, it has been trained on an image distribution that is very far from e-commerce, as it has never (or rarely) seen such pictures. Second, the set of classes it has been trained on is different from a set of classes that could be meaningful in e-commerce. A useful representation should separate different types of clothing (*e.g.* a skirt and a dress), but it should also discriminate between different lengths of sleeves for shirts, trouser cuts, types of handbags, textures, colors, shapes,...

Our goal is to learn a visual feature extractor designed for e-commerce images. This representation should:

- encode multiple levels of visual semantics: from low level signals (color, shapes, textures, fabric,...) to high

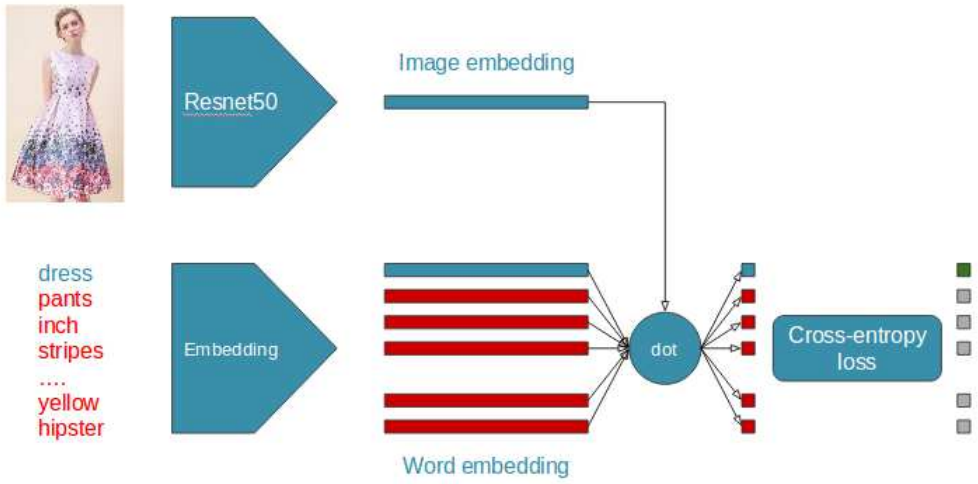

Figure 2. Training of our model: predict one label, picked from the bag-of-words description, from an image. Both image and words are embedded before being coupled in a dot product. We output a probability for each word in the vocabulary.

level information (style, brand),
- be separable over visual concepts, so we can train very simple classifiers over clothing types, colors, attributes, textures, *etc.*,
- provide a meaningful similarity between images, so we can use it in the context of image retrieval.

To these ends, we train a visual feature extractor on a large set of weakly annotated images crawled from the Internet. These annotations correspond to the textual description associated to the image. The model is learned on a dataset at *zero* labeling cost, and is exclusively constituted of data points extracted from e-commerce websites. Our main contribution is an in-depth analysis of the model presented in [9], through applications to fashion image recognition tasks such as image retrieval and attribute tagging. We also improved the method by upgrading the CNN architecture and dealt with multiple languages, mainly English and French. In Section 2, we explain the model, how we handle noise in the dataset, as well as some implementation details. In Section 3, we provide results given by our representation on image retrieval and classification, over multiple datasets. Finally in Section 4, we conclude and go over some possible improvement tracks.

2. Learning Image and Text Embeddings with Weak Supervision

One major issue in applied machine learning for fashion is the lack of large clothing e-commerce datasets with a rich, unique and clean labeling. Some very interesting work has been done on collecting datasets for fashion [11, 12]. However, we believe it is very hard to be exhaustive in describing every visual attribute (pieces of clothing, texture, color, shape, *etc.*) in an image. Moreover, even if this labeling work could be perfectly carried, it would come at very high cost, and should be manually done each time we wanted to add a new attribute. A possible source of annotated data is the e-commerce website catalogs. They provide a great amount of product images associated with descriptions, such as the one in Figure 1. While this description contains information about the visual concepts in the image, it also comes with a lot of noise that could harm the learning.

We explain now the approach we used to train a visual feature extractor on noisy weakly annotated data.

2.1. Weakly Supervised Approach

Learning with noisy labeled training data is not new to the machine learning and computer vision community [6, 17, 19]. Label noise in image classification [22] usually refers to errors in labeling, or to cases where the image does not belong to any of the classes, but mistakenly has one of the corresponding labels. In our setting, in addition to these types of noise, there are some labels in the classes vocabulary that are not relevant to any input. Text descriptions are noisy as they contain common words (*e.g.* 'we', 'present'), subjective words (*e.g.* 'wonderful') or non visual words (*e.g.* 'xl', 'cm'), which are not related to the input image. As we don't have any prior information on which labels are relevant and which are not, we keep the preprocessing of textual data as light as possible.

2.2. Model

Our work builds upon the one presented in [9], which we explain in this section. The model's training scheme is exposed in Figure 2

Figure 6. t-SNE map of 1,000 image samples from DeepFashion Categories dataset based on our Weakly image features extractor. We can identify some local subcategories, such as colorful dresses (a), black pants (b), stripes (c), checked (d) or printed shirt (e).

	bag	belt	body	bra	coat	combi	dress	eyewear	gloves	hat	neckwear	pants	shoes	shorts	skirt	socks	top	underpants
ImageNet	99.63	98.02	98.20	99.27	99.00	96.01	98.68	99.93	**99.99**	99.45	99.18	99.46	99.87	99.23	98.67	99.35	98.67	99.5
Weakly	**99.86**	**99.46**	**99.08**	**99.67**	**99.31**	**98.11**	**99.13**	**99.99**	99.97	**99.73**	**99.33**	**99.56**	**99.93**	**99.32**	**99.21**	**99.83**	**99.26**	**99.67**

Table 2. AUC classification score for clothing categories

	backpack	baguette	bowling bag	bucket bag	doctor bag	duffel bag	hobo bag	luggage	clutch	saddle bag	satchel	tote	trapeze
ImageNet	95.15	87.63	90.42	94.35	90.99	87.97	92.73	**87.65**	96.52	91.77	88.58	96.77	92.11
Weakly	**95.94**	**91.85**	**92.13**	**94.87**	**91.59**	**90.12**	**95.19**	86.96	**97.24**	**93.12**	**91.45**	**97.64**	**93.61**

Table 3. AUC classification score for fine-grained type of bags

Type dataset where images are annotated according to their clothing category (such as bags, shirt, dress, shoes, *etc.*). Table 3.3 shows the improvement on AUC scores over the ImageNet model for each of the clothing categories using our new representation. This indicates that our training scheme was able to learn discriminative features for garment classification.

Finally, we now focus on a fine-grained recognition task. The HandBag dataset contains images annotated with their specific type of bag. In this dataset, the differences between classes are more subtle than in the ClothingType dataset. The training and evaluation are the same as for the previous experiment. As in the previous experiment, we improved AUC scores for nearly each type of bags (see Table 3.3).

3.4. Exploratory visualization using t-SNE

To obtain some insight about our Weakly representation, we applied t-SNE [15] on features extracted using our Weakly feature extractor. We did this for 1,000 images from DeepFashion Categories test set. Figure 6 shows full map and some interesting close-ups. On top left (a), we can see a cluster of dresses sub-divised into multiple sub-clusters corresponding to different colors. The cluster (b) shows a focus on black pants. In the zone (c), we can easily see that the model gathered images containing stripes, and it seems like it has separated tops from dresses inside this cluster (with large striped sweaters on top). Checked clothings are grouped in cluster (d), while printed t-shirts are represented in cluster (e). This plot shows that our representation is able to group together concepts that are close in terms of clothing type, texture, color and style.

4. Discussion and Future Work

We presented in the future a method to learn a visual representation adapted to fashion. This method has the major advantage to overcome the issue of finding a large and clean e-commerce dataset. The results shows clear improvements compared to a visual representation trained on ImageNet, improving performance on multiple tasks such as image retrieval, classification and fine-grained recognition.

In the future, we would like to investigate on the possibility to better train our visual feature extractor using an external knowledge base of textual concepts.