Clothing Recognition in the Wild using the Amazon Catalog

Fabian Caba Heilbron
Amazon Berlin (Internship)

Bojan Pepik
Amazon Berlin

Zohar Barzelay
Amazon Berlin

Michael Donoser
Amazon Berlin

Abstract

The emergence of online influencers, the explosion of video content, and the massive amount of movie collections have served as an advertising vehicle for the fashion industry. This trend has created the need for automated methods that recognize people's outfit in such image and video collections. However, existing computer vision solutions for fashion recognition require an enormous amount of labeled data for training, which is prohibitively expensive. In this work, we propose an approach to build clothing recognition models for real-world scenarios. Our approach exploits images from the Amazon Catalog as training data. By using the catalog data as an additional training source, we boost the recognition accuracy on the challenging real world images of the DeepFashion dataset achieving state-of-the-art performance. We introduce the first dataset for clothing recognition in movies. In this scenario, we find that the use of catalog data for training becomes even more crucial, as it provides an accuracy boost of 10%.

1. Introduction

Imagine you are watching a movie and you like the outfit your favorite actor is wearing. How would you buy or find this outfit or other items in an online catalog? You would probably need to enter textual queries to limit your search, followed by spending hours navigating a series of sites and webpages before you can find something similar to the desired outfit. This search problem, combined with the fact that movies, TV shows, and media influencers are vehicles to dictate fashion trends and consumption, motivates online retailers to link their current inventory against real-world content. Doing this task manually would involve substantial expenses due to the large volume of data to be scanned by annotators. Thus, it becomes crucial to develop visual systems capable of automatically recognizing and indexing clothing items in the wild.

In this work, our aim is to automatically tag clothing items from images and movies in the wild. Specifically, our goal is to recognize clothing categories such as *Suit Jacket, Shirt, Jeans*, among others, relying primarily on

Figure 1. The goal of this work is to automatically tag the clothing item categories in real world images and movies.

catalog data. Unfortunately, fine-grained clothing category information is not always available in the catalog. Therefore, the focus of this work is to explore strategies to semi-automatically retrieve fashion category data from the Amazon catalog. Moreover, we study how this data can be used to improve performance for the challenging "in the wild" scenarios: real-world images and movies.

There are several attempts at solving fashion recognition by leveraging strong supervision. These models often rely on large-scale and manually curated datasets. For instance, [5] introduces DeepFashion, a large-scale dataset for fashion recognition, with the end goal of training convolution neural networks (CNN) to predict clothing attributes and categories. To build such database, they hire multiple annotators to go through each image, assign a clothing category, and draw a bounding box around the clothing item. Moreover, recent methods often require additional annotations such as clothing landmarks ([5], [7]). These annotations are unfeasible to obtain for data in the wild. We address these limitations and devise a method relying only on weakly supervised data, retrieved from the Amazon catalog using product metadata. We thus avoid the use of expensive annotations such as bounding boxes or clothing landmarks for training our models.

The contributions of this work are fourfold. First, we design a strategy to retrieve images from the Amazon catalog and build a dataset with weak labels extracted from product metadata. This strategy allows us to build a dataset of nearly 1 million images distributed among 43 fashion categories. Second, we build a simple yet effective CNN model to recognize clothing categories. It improves the state-of-the-art on public benchmarks, without using costly annota-

tions like clothing landmarks. Third, we conduct in-depth experiments to better understand how useful is the catalog data for clothing recognition in the wild. Specifically, we show the catalog data provides a rich source of data for rare categories. Moreover, we find that when the catalog data is combined with in-the-wild images for training, it can push further the accuracy by 2.1%. Finally, we introduce a novel dataset for clothing recognition in movies. The dataset contains more than 1000 one-second clips, augmented with clothing category information of actors. Experimental results in this new and challenging benchmark demonstrate the benefits of exploiting catalog data, especially when little data of the target domain is available.

2. Building Amazon Catalog Fashion Datasets

In this section we describe our strategy to build datasets for fashion recognition using the Amazon catalog. We achieve this with very little human intervention during the construction process.

2.1. Data Collection Pipeline

Our data collection process comprises of three steps. First, we define the categories and search queries to retrieve the images from the Amazon catalog. Then, we inspect the quality of the retrieved results and iterate over the queries if needed. Finally, we split the data into train and validation subsets. We describe below in detail each of these steps.

Our first task is to define a set of clothing categories. For the sake of simplicity, we clone the category set from the Deep Fashion [5] dataset. This would directly allow us to understand the effect of using catalog data to classify in-the-wild images. Next, we construct search queries to crawl the Amazon Catalog. We predominantly rely on the product metadata such as the title and the detail page text. In most cases, we apply simple filters such as making sure the category name appears in the product title and description.

After defining the set of products per category, next we select the category images and verify they align with the category definition. We retrieve all product images and manually prune images which do not show the actual product. To this end, we construct a tool for viewing a large sample of images at a quick glance. Note that unlike manually annotating every single image, with this tool we are only removing clear outliers, making the workload minimal and the process scalable. To make sure our train and validation sets do not overlap, we verify that a product does not appear across different sets. Further strategies like image deduplication can be applied but we leave it as future work.

2.2. Catalog Fashion Datasets

Using the data collection pipeline described above, we create a novel dataset for clothing recognition from Amazon catalog (AC) images. The AC dataset has 43 classes,

Dataset	#Images
Deep Fashion	140K
CatalogFashion	140K
CatalogFashion-10x	1M

Figure 2. (Left) Dataset statistics. (Right) Modes of the retrieved images. The first two modalities (from left-to-right) account for 95% of the downloaded images.

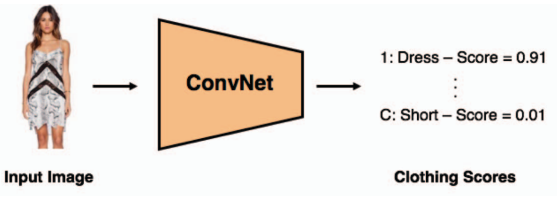

Figure 3. Our baseline model takes as input an image and produces a set of clothing scores.

which are identical to the 43 classes in the DeepFashion dataset [5]. The main difference between both datasets resides in the total number of samples. While [5] contains around 140K images, the CatalogFashion-10x has nearly one million. To analyze the effects of dataset size, we also have the CatalogFashion dataset with 140K images. Figure 2 (left) summarizes the datasets. The dataset contains sufficient number of samples per class, even for underrepresented categories (e.g. *Sarong* has around 1500 images).

We find three dominant data modes in the retrieved catalog data. First, a huge portion of retrieved images contain the clothing isolated (product only). In the second mode typically a person is wearing the clothing item (product wear on). Both modes can include images with background. In total they account for 95% of the retrieved images. The third modality, which we scarcely saw, include images with the clothing item in the wild. We present examples of each modality in Figure 2 (Right).

3. Clothing Recognition Model

Our goal is to train a model to recognize clothing items in images. Given an input image, the model generates per-class output scores, which associate the chance of each clothing category appearing in the picture. To that end, we train a CNN in a supervised fashion by relying on a dataset containing pairs of images and clothing labels. We optimize the Cross Entropy loss [2], which is widely used for multi-class recognition problems. For a given input image (or mini-batch of images), the loss computes the distance between the output probability distribution over the clothing classes, and the ground-truth labels. Once the model has been trained, it can predict scores for the clothing categories from the training dataset (See Figure 3).

In terms of implementation details we use ResNet-50 V2 [4] as the backbone architecture. We exploit available

Train Source	Top 3 Acc	Top 5 Acc
DF [5]	87.8%	93.3%
CatalogFashion	71.0%	80.4%
CatalogFashion-10x	76.5%	84.4%
CatalogFashion+DF	88.2%	93.6%
CatalogFashion-10x+DF	**89.9%**	**94.7%**

Table 1. Results on the DeepFashion [5] validation dataset.

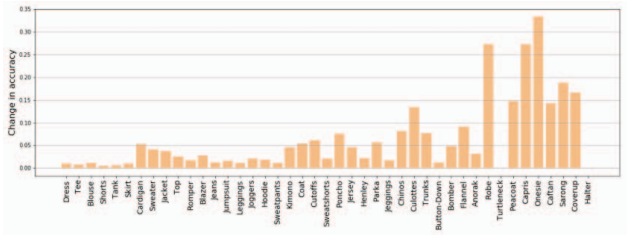

Figure 4. Top-3 accuracy difference between the models trained on catalogFashion-10x+DF and DF. Differences across categories, sorted in decreasing order of number of samples.

pre-trained weights from ImageNet [6] for initialization. To train we rely on Stochastic Gradient Descent (SGD) as optimizer with momentum 0.9 and initial learning rate 10^{-1}. We apply learning rate decay every 25 epochs by a factor of 0.1. For data augmentation we apply: random cropping, randomly flipping the image from left to right, and random color jittering. Importantly, all images are: (a) re-scaled such the smaller axis dimension is equal to 256 while preserving the original aspect ratio, and (b) normalized with ImageNet RGB mean and standard deviation values. Our implementation relies on MXNet Gluon [1].

4. Experiments

4.1. Clothing Recognition For Images in the Wild

We start the experimental evaluation by focusing first on images. Our goal is to investigate to what extent we can use catalog data in the real world. In a nutshell, we found that the catalog data helps to train better models even when there is a large domain shift.

Experimental setup. For training purposes we mainly utilize the datasets introduced in this work. In some of the experiments, we also use the training set of the DeepFashion dataset. For evaluation we use the DeepFashion validation subset (40K images). With the end goal of conducting experiments in a close to real-world scenario, where bounding boxes are costly to obtain, we train all models using image-level annotations only (unless otherwise mentioned). Thus, we discard bounding boxes and fashion landmarks existing in the DeepFashion dataset. Similarly, at test time we do not use BBs or landmarks.

In terms of evaluation, the Deep Fashion authors propose to use top-3 and top-5 accuracy in order to cope with label noise. We follow their evaluation setup.

Results. We start by evaluating the performance of our clothing recognition model on DeepFashion. Table 1 reports the results. First, we would like to understand if the catalog can be used to recognize clothing categories. We observe that the CatalogFashion and CatalogFashion-10x datasets with 71.0% and 76.5% top-3 accuracy show competitive performance to the model trained on DeepFashion data (87.8%). We conclude that catalog data, despite the gap in the data distribution, can still be resourceful for recognizing clothing categories in real imagery. Second, we observe that the more data from the catalog we use, the better the performance: the CatalogFashion-10x dataset outperforms the CatalogFashion one by 5.5% top-3 accuracy.

Next, we are interested to see if the Amazon Catalog can be leveraged to improve the overall clothing recognition performance on the dataset. To that end, we combine the AC datasets with the DeepFashion one for training. Table 1 demonstrates that the AC data can help in improving the overall performance. For both AC datasets we observe improvements when adding the DeepFashion dataset, and with 89.9% top-3 accuracy CatalogFashion-10x combined with the DeepFashion dataset sets the highest performance.

Next we would like to understand which classes does the catalog data help the most. We report in Figure 4 the accuracy change per class between a model trained with the catalogFashion-10x+DF data and a model trained with the DF data. The higher the bar, the more the catalog data helps in recognizing that class. Interestingly, we improve accuracy for almost all classes. Note that classes are sorted (decreasing order) based on the number of samples in the DF train set. The catalog data becomes crucial when few samples are available in the target dataset. In addition, we analyzed the performance gains across different data types: upper, lower and full body. Interestingly, adding the catalog data results in higher improvements for full-body items.

State-of-the-art comparison. Finally, we compare our models to the state-of-the-art. Table 2 compares our model results against previous works [5, 7], which rely on landmark and bounding box annotations. We observe that our model outperforms the FashionNet approach even without using bounding boxes at test time, and with 91.4% outperforms the previous state-of-the-art. Note that at training time our method only uses image-level annotations.

4.2. Clothing Recognition in Movies

Encouraged by the results on images, we now proceed with clothing recognition in movies. Unlike the constrained

Method	Top 3 Acc	Top 5 Acc
FashionNet [5]	82.6%	90.2%
Grammar Networks [7]	91.0%	95.8%
Our	89.9%	94.7%
Our + BB	**91.4%**	**95.8%**

Table 2. State-of-the-art comparison on DeepFashion [5]. Our model improves the state-of-the-art without using landmarks and achieves competitive performance without using bounding boxes.

Training Source	Top 3 Acc	Top 5 Acc
AVA-Fashion	43.1%	56.9%
AVA-Fashion+CF-10x	**53.1%**	**65.5%**

Table 3. Clothing recognition performance on AVA-Fashion.

scenarios of people facing the camera to pose for the picture, recognizing clothing items in movies is very challenging. Partial views, low resolution, challenging poses are among the challenges specific to movie data. To the best of our knowledge there are no previous studies that investigate this task in movies. In this work, we make a step closer and build a dataset for fashion recognition in movies.

AVA-Fashion - a Novel Dataset for Fashion Recognition in Movies. We collect a new dataset that contains more than 1k one-second video clips with a bounding box localizing an actor. We exploit the existing BBs from the AVA dataset [3]. We augment the BBs with clothing categories. We sampled 4k annotations from AVA and manually assign categories from DeepFashion. A huge portion of the AVA annotations only displays a person's face, show clothing items not within our lexicon, or the bounding boxes cover very few pixels; in those cases we discard the clips. In fact, we were able to annotate 25% from the sampled clips with the fashion categories in our vocabulary.

AVA-Fashion experiments. We first split the collected annotations into training and validation. We use the training data to fine-tune our models, and report top-3 and top-5 accuracies on the the validation subset.

Our model follows the architecture defined in Section 3. The analyzed models differ in terms of training data. Specifically, we study again the effect of using catalog data for training. To that end, we compare using AVA-Fashion for training only versus using the AC data.

Table 3 summarizes the results. Catalog pre-training offers a 10% boost in top-3 accuracy w.r.t. models trained on Ava-Fashion, reiterating the benefits of catalog images, even under large domain shifts. Our best model achieves an encouraging top-5 accuracy of 53.1%. Although there is space for improvements we believe we made initial baby steps towards clothing recognition in the movie domain.

5. Conclusions

We studied the effectiveness of using images from the Amazon catalog to train clothing recognition models. Our experiments were conducted in two challenging scenarios: consumer photos and movies. We investigated using Amazon catalog data for clothing recognition in the wild. Our findings reveal that the catalog data is essential when there is relatively little available data in the target domain. We opened new venues of research and development by constructing the AVA-Fashion dataset. It is a dataset for fashion recognition in movies, which is the first of it kind, and presents a challenge for existing image-based solutions.

References

[1] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library, 2015. In NIPS, 2016. 3

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 2

[3] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2, 3, 4

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 3

[7] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 1, 3, 4