Multi-Label Fashion Image Classification with Minimal Human Supervision

Naoto Inoue$^1$  Edgar Simo-Serra$^2$  Toshihiko Yamasaki$^1$  Hiroshi Ishikawa$^2$

$^1$The University of Tokyo  $^2$Waseda University

inoue@hal.t.u-tokyo.ac.jp  esimo@aoni.waseda.jp  yamasaki@ay-lab.org  hfs@waseda.jp

Abstract

We tackle the problem of multi-label classification of fashion images, learning from noisy data with minimal human supervision. We present a new dataset of full body poses, each with a set of 66 binary labels corresponding to the information about the garments worn in the image obtained in an automatic manner. As the automatically-collected labels contain significant noise, we manually correct the labels for a small subset of the data, and use these correct labels for further training and evaluation. We build upon a recent approach that both cleans the noisy labels and learns to classify, and introduce simple changes that can significantly improve the performance.

1. Introduction

Fashion plays an important role in everyday lifestyle, yet understanding fashion is still a very complicated task for computer vision. In particular, due to the subjectiveness of the fashion, obtaining high-quality data for training learning-based models for fashion task remains an open problem. In this work, we analyze the importance of the quantity of the data and evaluate the modern tools for automatically cleaning the data in the context of multi-label prediction.

Unlike more standard computer vision tasks, the fashion domain is variable and subjective, and networks pre-trained on large-scale datasets such as ImageNet and Places are not necessarily best-suited for fashion-related tasks. Simo-Serra and Ishikawa [21] introduced Fashion144k, which consists of fashion images in a variety of scenes with weak labels. Furthermore, they showed that exploiting a large number of fashion images with only weak labels and designing a CNN to learn from them can provide a better feature representation than ImageNet or Places for one of the fashion-related tasks, i.e., fashion style classification. Thus, obtaining a larger-scale fashion image dataset is desirable as a basis for fashion recognition tasks. We examine two aspects that have an impact on the recognition performance on fashion image datasets: (i) the size of the dataset and (ii) the quality of the annotations in the dataset. We first extend the Fashion144k dataset [20] by crawling over 1M images with metadata, which contains weak labels, from Chictopia$^1$. To evaluate the recognition performance, we perform an experiment in multi-label classification, as shown in Fig. 1, and predict the colors and the garments worn by the person in each image. To evaluate the quality of annotations, we randomly pick 5,300 images in our new dataset, which we denote as Fashion550k, and manually correct the weak labels to obtain “clean” labels. We use these limited annotations to both evaluate and improve the recognition accuracy. Note that the number of the labels in this subset is quite small, compared to the whole dataset. We build upon the approach of Veit et al. [25] that uses a neural network to clean the labels by learning a mapping between the noisy labels and clean labels. By learning this mapping, it is possible to afterwards jointly clean the dataset and train a prediction model. We evaluate this approach and show that with some simple modifications we are able to significantly improve performance.

To summarize, our contributions are as follows:

---

$^1$Dataset available at http://hi.cs.waseda.ac.jp/~esimo/data/fashion550k/.
• We present a large-scale weakly-labeled fashion image dataset that includes 5,300 human-annotated images for analyzing the label noise.
• We investigate the effect of the size of the training data for multi-label fashion classification.
• We evaluate the methods to exploit weak labels jointly with human-verified data.

2. Related Work

2.1. Fashion

There are many fashion-related problems in computer vision research, such as semantic segmentation of garments [19, 29, 31, 13, 12], image retrieval [8, 5, 33], fine-grained classification of garments and styles [2, 3, 14, 10, 30], fashion landmark detection [15], and image-conditional domain transfer [32]. Predicting more subjective properties such as popularity [26] and fashionability [20] have also been studied. Many of these approaches collected large scale image datasets by crawling the Internet. They also exploit weak labels [20, 28] or manually annotate the images by crowd-sourcing [14]. In this paper, we construct the Fashion550k dataset that consists of fashion images with weak annotations, and evaluate the influence of noisy annotations. We believe this dataset is also useful for other tasks such as fashionability estimation and clothing parsing (as a weak supervision).

2.2. Learning from Noisy Data

There are two approaches to learn from noisy labeled data with a neural network: the first approach aims to directly learn from noisy labels and focus mainly on constructing a model that takes noise into consideration. The simplest approach is to model a label noise that is conditionally independent of the input image [18, 23]. Xiao et al. [27] and Misra et al. [17] proposed image-conditioned noise model. However, these methods face the challenge of distinguishing difficult samples from mislabeled training samples, and rely on heuristic-based noise models.

Another approach is called semi-supervised approach, which is to learn from noisy labeled data with neural network is to combine a small set of clean labels. Veit et al. [25] proposed to learn the mapping between noisy and clean labels and then to exploit the mapping for training deep neural networks by imitating the student-teacher models used in [16, 7]. In this approach, a base CNN is used as a feature extractor and is combined with two additional networks consisting of a label cleaning network and an image classifier. The label cleaning network is input the noisy labels in addition to the visual features, and is trained to predict cleaned labels that are provided as a privileged information. The output of the cleaning network is regarded as the target label to predict for the image classifier. This approach can handle both incorrect and missing labels without assuming the type of noise and greatly boost the classification performance. We evaluate this approach, propose simple changes to improve the performance, and furthermore propose an alternative scheme in which the confidence of the label cleaning network is used.

3. Dataset

We extended the Fashion144k dataset [20]. The extended dataset, called Fashion550k, consists of 550,661 user posts from the clothing-oriented website Chictopia. Fashion550k is about four times the size of the original Fashion144k. In Fig. 2, example images of Fashion550k are shown. Each post contains at least a single image with metadata including label tags. Each photograph shows a different angle of the user or a zoom-in on different garments. Users often add a short description of what they are wearing and tags of the types and colors of the garments. Note that this information is very noisy: e.g., not all garments are tagged, and some users make only a part of the information available. We parsed all the information provided with each post. For a representative statistics of the dataset, refer to Table 1.

Since Fashion550k has been collected without any filtering, some of the images are not suitable for learning. For example, some images focus on only one object, or have strong filters. Thus, we cleaned the data following the ap-
Table 1: Statistics of the dataset. The number between the brackets indicates standard deviation of each property.

<table>
<thead>
<tr>
<th>Property</th>
<th>Total (Unique)</th>
<th>Per user</th>
<th>Per post</th>
</tr>
</thead>
<tbody>
<tr>
<td>posts</td>
<td>550,661</td>
<td>36.19 (71.41)</td>
<td>-</td>
</tr>
<tr>
<td>users</td>
<td>15,217</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>photos</td>
<td>1,061,468</td>
<td>1.76 (0.90)</td>
<td>1.93 (1.26)</td>
</tr>
<tr>
<td>tags</td>
<td>3,627</td>
<td>1.98 (1.55)</td>
<td>2.14 (2.03)</td>
</tr>
</tbody>
</table>

As weak annotations, we utilize tags that consist of colour-garment pairs, such as blue-jeans and red-sweater. We split the tags into colors and garments, resulting in a total of 123 unique weakly-annotated tags. However, some tags, such as orange and carrot-orange, or watch and bracelet, are hard to visually discriminate under various illumination and lighting conditions. Furthermore, some tags are hard to detect because occurrence is minimal or because they are highly occluded, as in the case of iPhone-case and earrings. Thus, we discard or merge such tags, resulting in a total of 66 unique weakly-annotated tags. The classes in this dataset are not evenly distributed, as shown in Fig. 3a. The class black has more than 200,000 annotations, whereas the class tie has only 1,435 annotations.

We additionally verified the tags in a subset of the collected images to provide minimal supervision. The number of the images verified was 3,000 for training, 300 for validation, and 2,000 for testing. The rest of the images (called the noisy dataset) are used to train the baseline classifier. We estimate the quality of the noisy labels using the verified tags for the 5,300 images. Fig. 3b shows the distribution of the quality of the original noisy labels. The noise occurs regardless of the class frequency. We observe that 26.4% of the original labels are false positives. Further, we also observe that originally positive labels accounts for only 54.2% of the labels actually verified positive. These observations show that directly using these weak labels as ground truth will lead to poor performance, which we will later validate in our experiments.

4. Proposed Method

In this paper, our goal is to train a multi-label image classifier on a large dataset with extremely noisy labels, where additionally a small subset of the dataset that has human-verified labels available. This setting can often happen when we collect images from the web or social media and have experts to correct some of the labels.

Formally, we have a very large training dataset $T$. $T$ consists of tuples of noisy labels $y$ and images $x$, $T = \{(x_i, y_i), \ldots\}$. Additionally, we have a small dataset $V$ with human verified labels $v$, $V = \{(x_j, y_j, v_j), \ldots\}$. The number of the data in $T$ is significantly larger than that in $V$. In our experiments, $T$ exceeds $V$ in the number of data by two orders of magnitude. Each $y$ and $v$ is a sparse $d$-dimensional vector with a binary annotation for each of the $d$ classes indicating whether it is present in the image or not. Our aim is to fully utilize the accurate annotation verified by a human in $V$ and noisy but huge number of labels in $T$.

4.1. Multi-Task Label Cleaning Network

We base our approach on the model of Veit et al. [25], in which a label cleaning network is used in combination with a classification network. We next briefly summarize [25].

4.1.1 Model architecture

The network is designed to jointly learn to generate accurate labels from noisy labels and to learn a more accurate multi-label classifier from the generated labels. An overview of the model is shown in Fig. 4. There are two classifiers $g$ and $h$ on top of a CNN-based feature extractor $f$.

The first classifier $g$, shown on the bottom of Fig. 4, is called the label cleaning network. It learns a mapping from noisy labels $y$ to human-verified labels $v$, conditional on the input image. Its output $c$ denotes the cleaned labels. The classifier $g$ has two separate inputs, the noisy labels $y$ and the visual features $f(x)$. Each input is projected into an embedding by a linear layer and the two are concatenated, then transformed with a hidden linear layer. Finally, $y$ is added to the output by an identity-skip connection and clipped to $[0, 1]$ to remain in the valid label space. In short, $c$ is com-

Figure 3: Label statistics for Fashion550k. We show and order by the class frequency, and represent the quality of the original labels as the accuracy of the noisy labels compared to manually cleaned labels.
Two losses are used to train the model: the label cleaning loss \( L_{\text{clean}} \) to enhance the quality of the cleaned labels \( c \) and the classification loss \( L_{\text{classify}} \) to enhance the quality of the predicted labels \( p \).

The label cleaning network is supervised by the verified labels of all samples \( v \) in the human verified set \( V \). The cleaning loss is based on the \( L_1 \)-distance between the cleaned labels \( y_j \) and the verified labels \( v_j \).

\[
L_{\text{clean}} = \sum_{j \in V} |c_j - v_j| \tag{2}
\]

The classification network is also supervised by two terms. For all samples \( i \) from the noisy subset \( T \), the image classifier is supervised by the cleaned labels \( c_i \) that is produced by the cleaning network. For samples \( j \) from verified dataset \( V \), we can directly supervise \( p_i \) by verified labels \( v_i \). For both terms, a cross-entropy loss is used.

\[
L_{\text{classify}} = -\sum_{j \in V} p_j \log(v_j) - \sum_{i \in T} p_i \log(c_i) \tag{3}
\]
sion ($mAP$) are used. $AP_{\text{all}}$ regards every annotation for all classes equally by handling them as coming from one single class and thus is biased towards more frequent classes.

5.1. Baseline and Compared Methods

As the baseline model for our evaluation, we train a CNN on the noisy labels from Fashion144k and Fashion550k. 117,746 and 402,472 images are used to train on Fashion144k and Fashion550k, respectively. This model is called Baseline and is used as the initial weight for all the other compared methods except ours. The Baseline model is based on a 50-layer ResNet [6] pre-trained on ImageNet [4]. The last softmax layer of the network is replaced with a 66-way sigmoid layer to predict the probability of each of the labels.

We compare the following approaches:

**Fine-tune with clean labels.** Most common approach with cleaned labels is to feed the clean labels directly to the network and supervise the last layer. However, this approach is prone to overfitting, since the number of clean labels is limited and cannot utilize huge noisy labels fully.

**Fine-tune with mix of clean and noisy labels.** This approach handles the insufficient training examples. We fine-tune the last layer of the base CNN with a combination of training samples from the small clean and the large noisy subset (in a 1 to 9 ratio).

**Veit et al. [25].** This approach utilizes the clean labels using the label cleaning network. There are two variants of this method, “with pre-training” and “trained jointly”. In “with pre-training”, first, we train just the label cleaning network. Subsequently, the classification network and the image classifier are jointly trained. In this training, a learning rate of the classification network is smaller ($\times 10^{-1}$) than the image classifier. In “trained jointly”, the classification network and the image classifier are trained jointly from the beginning. A difference with the original implementation is that in this case we opt to use the ResNet50 [6] as the Base CNN instead of the Inception v3 model [24]. For further details, please refer to the original paper.

**Improved Model.** This approach only implements our modification of adding ReLU and BN to the label cleaning network, which is proposed in Section 4.2. This model is trained in the same way as original method of Veit et al.

**Ours.** This approach implements both using Improved Model and the two-phase training shown in Fig. 5b and Fig. 5c. To reduce the computational overhead, we initialize the base-CNN and the prediction network with the weight of Fine-tune with clean labels.

5.2. Training Details

All approaches are implemented using PyTorch [1] and optimized with Adam [11] by a batch size of 64. All the images are resized to 256 x 256. In the training phase, the images are then randomly cropped into 224 x 224 with random horizontal flipping. In the test phase, the images are center-cropped into 224 x 224. We employed early stopping using $AP_{\text{all}}$ on the validation subset.

The baseline network is initially trained with the binary cross-entropy loss between the noisy labels and the predictions of the network for 100,000 iterations using an initial learning rate of $10^{-5}$. For fine-tuning, we use a learning rate of $10^{-4}$ for the last linear layer and $10^{-5}$ for the other layers, for additional 20,000 iterations.

The other variants are trained for additional 20,000 iterations. In this paper, we regard the ResNet except the last two layers, i.e., linear layer and sigmoid layer, as the base-CNN. The last two layers are regarded as the image classifier. For the label cleaning network, the arrangement of the linear layers are the one used in Veit et al. as in Fig. 4 and the number of filters for the each linear layer is set to 512. We use an initial learning rate of $10^{-5}$ for the base-CNN and $10^{-4}$ for the other layers. To train the cleaning network and image classifier jointly, we sample training batches that contain samples from $T$ as well as $V$ in a ratio of 9:1. To balance the losses, we weight $L_{\text{clean}}$ with 0.1 and $L_{\text{classify}}$ with 1.0 for the variants of Veit et al., which is the same as the parameters used in the paper.

5.3. Results

We first discuss the overall performance of the proposed method in Table 2. The performance regarding $AP_{\text{all}}$ is higher than $mAP$. This means that the $AP$ for popular classes is higher. Training solely on the noisy labels from our Fashion550k, which is about four times bigger than Fashion144k [20], shows significant performance gain and shows about +7% improvement in $AP_{\text{all}}$ and about +9% improvement in $mAP$ compared to training solely on noisy labels from Fashion144k. This result suggests that collecting larger dataset is still necessary for the recognition performance improvement.

<table>
<thead>
<tr>
<th>Model</th>
<th>$AP_{\text{all}}$</th>
<th>$mAP$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Fashion144k)</td>
<td>62.23</td>
<td>49.66</td>
</tr>
<tr>
<td>Baseline (Fashion550k)</td>
<td>69.18</td>
<td>58.68</td>
</tr>
<tr>
<td>Fine-tuning with mixed labels</td>
<td>72.38</td>
<td>61.50</td>
</tr>
<tr>
<td>Fine-tuning with clean labels</td>
<td>79.39</td>
<td>64.04</td>
</tr>
<tr>
<td>Veit et al. (pretrain)</td>
<td>78.60</td>
<td>62.81</td>
</tr>
<tr>
<td>Veit et al. (joint)</td>
<td>78.92</td>
<td>63.08</td>
</tr>
<tr>
<td>Improved Model (pretrain)</td>
<td>80.01</td>
<td>64.34</td>
</tr>
<tr>
<td>Improved Model (joint)</td>
<td>79.70</td>
<td>64.03</td>
</tr>
<tr>
<td>Ours</td>
<td>79.87</td>
<td>64.62</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the models evaluated by the test subset of Fashion550k. We use all 3000 images with cleaned labels for the training. (noisy:clean ≃ 136:1)
Table 3: Examples from the test subset in Fashion550k. We show the top 5 most confident predictions along with whether the prediction is correct or incorrect. Our approach consistently removes false predictions.

<table>
<thead>
<tr>
<th>Image from the test subset</th>
<th>Top5 predictions</th>
<th>Baseline</th>
<th>Fine-tuning with clean labels</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dress</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bag &amp; Purse</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sunglasses</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>White</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pink</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sandals</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Skirt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Black Stockings</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>&amp; Tights</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Brown</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Red</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Shirt</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Ruby-red</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Shoes</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
</tbody>
</table>

Then we conduct an analysis on combining noisy and clean labels. Simply fine-tuning with clean labels improves $AP_{all}$ and $mAP$ significantly because Fashion550k is very noisy. We would like to emphasize that the two variants of the original Veit et al. do not work and are even worse than fine-tuning with clean labels. Further, there is still certain improvement in improved model and ours in both $AP_{all}$ and $mAP$ over fine-tuning with clean labels. This shows the importance of carefully designing the label cleaning network, which results in over 1% improvement both in $AP_{all}$ and $mAP$ compared to the original method of Veit et al. Notably, ours achieved best $mAP$. The difference in improved models and ours are not so remarkable. This suggests that the order of the training is less important. In Table 3, the example results of classification are shown.

We perform an analysis on a harder setting, where we only have access to only 1,000 cleaned annotations. The gap between fine-tuning with clean labels and the label cleaning network based methods are clear as shown in Table 4. Notably, the performance improvement by the proposed methods, i.e., +0.6% $\sim$ +0.8% in $AP_{all}$ is not subtle compared to the improvement when we use 3 times larger number of cleaned labels, i.e., +1.4% in $AP_{all}$, as shown in Table 2.

We perform another analysis, where we only have access to Fashion144k for noisy labels in Table 5. Here, $AP_{all}$ and $mAP$ get worse compared to fine-tuning with clean labels in almost all proposed methods. This result suggests that the label cleaning network-based approach does not work if we have relatively enough cleaned annotations.

### 6. Conclusion

We have presented a new large-scale weakly-labelled dataset for multi-label classification of garments of full pose images, providing a small subset of validated annotations for evaluation. We provide experimental evaluation of the effect of the number of images and number of clean annotations, along with various variants of multi-task classification networks that make use of the clean and noisy annotations. Results show that a few simple modifications are able to improve the performance of previous approaches significantly, and larger amounts of noisy data are useful for improving classification results.

**Acknowledgements:** This research was supported by GCL program of The Univ. of Tokyo by JSPS. N. Inoue and T. Yamasaki are partially supported by the Grants-in-Aid for Scientific Research (no. 26700008) from JSPS, JST-CREST (JPMJCR1686), and Microsoft IJARC core13.
Simo-Serra and H. Ishikawa are partially supported by JST-CREST Grant Number JPMJCR14D1, JST ACT-I Grant Number JPMJPR16UD.

References