The iMaterialist Fashion Attribute Dataset

Sheng Guo[1] Weilin Huang[1] Xiao Zhang[2] Prasanna Srikhanta[3] Yin Cui[4] Yuan Li[5]
Hartwig Adam[2] Matthew R. Scott[1] Serge Belongie[4]
[1]Malong Technologies [2]Google AI [3]Wish [4]Cornell University [5]Horizon Robotics

Abstract

Many large-scale image databases such as ImageNet were constructed only for single-label and coarse object-level classification, while multiple labels and fine-grained categories are often needed in real-world applications, yet very few such datasets exist publicly. In this work, we contribute to the community a new dataset called iMaterialist Fashion Attribute (iFashion) to address this problem in the fashion domain. The dataset was constructed from over one million fashion images with a label space that includes 8 groups of 228 fine-grained attributes in total. The result is the first known million-scale multi-label and fine-grained image dataset. We conduct experiments and provide baseline with various CNN models. Importantly, we demonstrate models pre-trained on iFashion can achieve better transfer learning performance on fashion-related tasks than ImageNet or other fashion datasets. Data is available at: https://github.com/visipedia/imat_fashion_comp.

1. Introduction

Recent deep learning models trained on large-scale datasets have significantly advanced various computer vision tasks, and the performance on existing image classification benchmarks such as ImageNet [2] has reached the saturation point [5, 15, 6]. New datasets need to be created to tackle more challenging problems, such as multi-label classification and fine-grained recognition. On the other hand, domain-specific datasets have raised a lot of interest, especially in fashion domain [18, 12, 4]. In light of this, we introduce an iMaterialist Fashion Attribute Dataset (iFashion), which includes over one million annotated fashion images where the labels are curated by fashion experts. The label space includes 8 groups and a total of 228 fashion attributes, as described in Table 1.

iFashion presents a few unique challenges. Firstly, it is a multi-label prediction problem and the models are evaluated by precision and recall. Most existing datasets created for multi-label image recognition are limited in scale, such as PASCAL VOC [3], COCO [11] and NUS-WIDE [1], which have about 6K, 80K and 160K training images from 20, 80 and 81 categories, respectively. Both learning difficulty and annotation effort would be increased considerably when the number of categories increases.

Secondly, many fashion attributes in iFashion are fine-grained labels and have very similar visual patterns. For example, as shown in Fig. 1, in the group of $Neckline$, identifying fine-grained visual difference on the defined fashion pattern ($Neckline$) between classes of U-$necks$ and $Shoulder$ is particularly challenging because the images often have large visual diversity within each class. This is much more significant than the subtle distinctions between different classes on the defined fashion pattern, resulting in significantly larger intra-class diversity than inter-class variance. This gives rise to new challenges compared to existing benchmarks for fine-grained recognition where images often have similar visual appearance with low intra-class diversity, such as CUB-200-2011 database [16] and Stanford Cars database [9]. More details on related studies with full comparisons between existing datasets are presented in the supplementary material.

The goal of iFashion is to encourage research on a more complex task toward real-world applications, by jointly considering multi-label and fine-grained image recognition with a hierarchical label structure. Our major contributions are: (i) the first known million-scale image dataset with multiple fine-grained attribute labels curated by experts; (ii) extensive experiments were conducted by using recent CNN models for multi-label and fine-grained recognition tasks, providing meaningful baseline results; (iii) we demonstrate empirically that iFashion is valuable for transfer learning on other fashion related datasets and applications.

2. iFashion Dataset

We describe the details of iFashion database. All images in iFashion are provided by Wish. We collected 1M+ fashion images by randomly sampling across individual attribute classes. All the images were pre-tagged by humans using an organically grown taxonomy. Then these tags were mapped to our taxonomy. Please refer to the supplementary material for post-processing steps we applied to improve

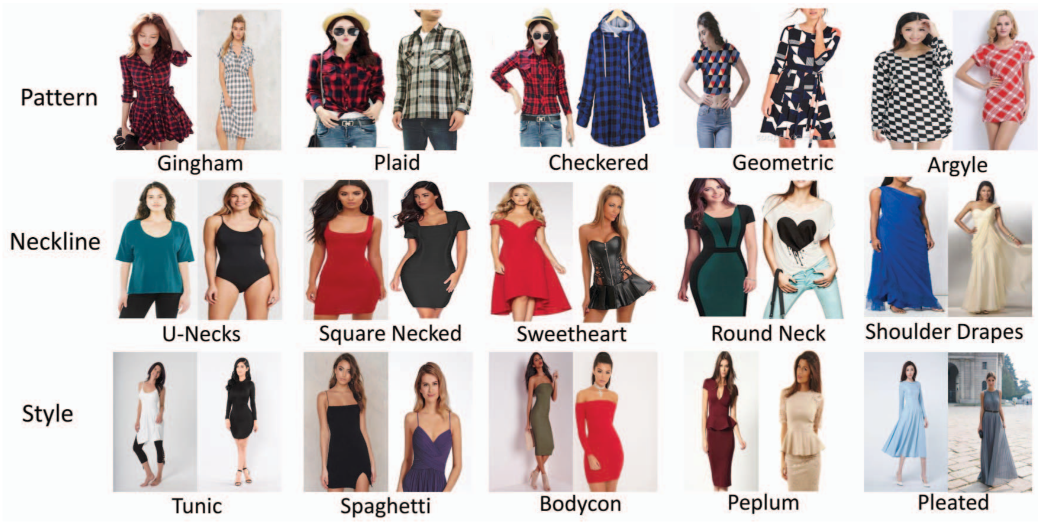

Figure 1. Examples from iFashion dataset for the attribute groups of $Pattern, Neckline, Style$.

Attribute	# Class	Type	# Label	# Image	Example
Category	105	S	913,857	913,857 - 90.2%	Athletic Pants, Bikinis, Cargo Pants, Heels, Petticoats ...
Color	21	M	894,904	467,137 - 46.1%	Black, Bronze, Gold, Gray, Green ...
Gender	3	M	1,012,947	935,265 - 92.3%	Male, Female, Neutral.
Material	34	M	701,197	591,175 -58.4%	Nylon, Organze, Patent, Plush, Rayon ...
Neckline	11	S	721,908	721,908 -71.3%	Racerback, Shoulder Drapes, Square Necked, Turtlenecks, U-Necks ...
Pattern	28	M	325,361	311,676 - 30.8%	Argyle, Camouflage, Checkered, Floral, Galaxy ...
Sleeve	5	S	733,501	733,501 - 72.4%	Long Sleeved, Puff Sleeves, Short Sleeves, Sleeveless, Strapless.
Style	21	S	610,442	610,443 - 60.3%	Asymmetric, Summer, Tunic, Vintage Retro, Wrap ...

Table 1. The number of classes, type (single-label or multi-label), number of labels and images for each attribute group in iFashion.

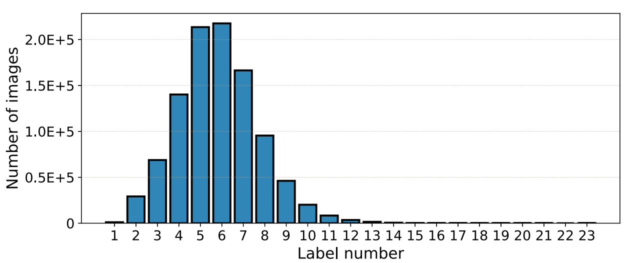

Figure 2. Histogram of number of labels per image, with an average of 5.8 and 8 per image in the training and validation sets.

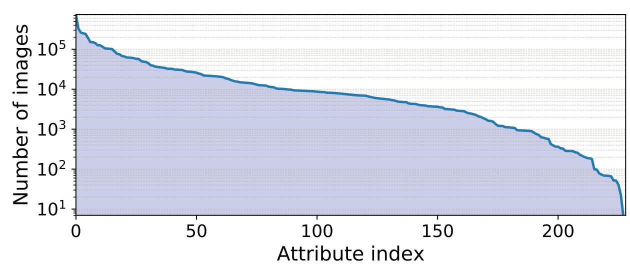

Figure 3. Number of images per attribute label, demonstrating the long tail nature of the dataset.

dataset quality. This results in iFashion database having 228 fine-grained attribute-level classes form 8 high-level groups defined professionally from the fashion industry. It contains 1,012,947 images for training, 9,897 and 39,706 manually-cleaned images for validation and testing.

Dataset statistics. The numbers of images and labels provided for each group are listed in Table 1. As can be found, the "Gender" group has a label in 92.3% images of the training set, while the "Pattern" group just has labels in 30.8% images. Histogram for the number of labels per image is shown in Fig. 2, where the number of labels per image is ranged from 1 to 23, with an average of 5.8. Furthermore, the number of images per attribute-level class is shown in Fig. 3, where 31 classes have $<$500 training images, while 88 classes are with $>$10K images, indicating significant data imbalance. In addition, recognition difficulty is changed significantly over different high-level groups or attribute-level classes. Fig. 4 shows top-8 recalls for attribute-level classes, and the average top-1 recalls for 8 groups correspond to Table 1 are: 58.5%, 48.3%, 97.3%, 52.2%, 66.0%, 43.1%, 86.2%, and 28.8%.

Our database considers large scale (million level), multiple labels (with group structure), and fine-grained recognition jointly for fashion recognition, setting it apart from

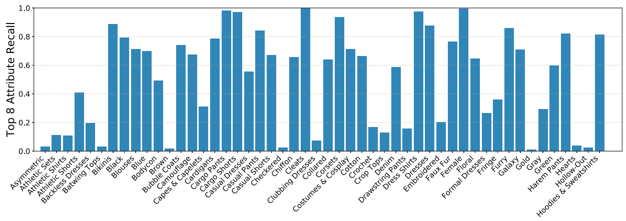

Figure 4. Per-attribute recalls on 50 randomly selected attributes.

existing datasets which were often designed for investigating each individual problems. In particular, our fine-grained classes are created structurally based on multiple groups of fashion attributes, which are professionally defined, such as "pattern", "neckline" and "color", as shown in Fig. 1. This allows it to have significantly larger intra-class variance than inter-class diversity, making it more challenging than existing fine-grained or fashion-related benchmarks. Our database sets a new challenge for CNNs to learn fine-grained distinctions between such structurally-defined patterns automatically from the provided data and labels.

Evaluation Metrics. Inspired by Wu *et al*. [17], we employ micro recall, micro precision and mean-F1 for measuring multi-label recognition. The "micro" means it is a measure over all images. We select top-8 output scores as the predicted results for each image as there are 8 labels on average for each image in the validation set. The evaluation metrics are computed via precision (P) and recall (R):

$$P = \frac{\sum_i^C N_i^t}{\sum_i^C N_i^p}, \quad R = \frac{\sum_i^C N_i^t}{\sum_i^C N_i^g}, \quad F1 = \frac{2 \times P \times R}{P + R}, \tag{1}$$

where C is the number of classes, N_i^t is the number of images correctly predicted for the i-th class. N_i^p and N_i^g are the numbers of predicted images and ground truth images. We apply the cross-entropy loss as

$$L(P, Q) = -\sum_i^C p_i \log q_i + (1 - p_i) \log (1 - q_i), \tag{2}$$

where C is the number of classes, P and Q denote a ground truth binary vector and the predicted probability scores.

3. Experiments, Baselines and Comparisons

We conduct experiments by using recently proposed CNN models, and provide baseline results with discussions. Furthermore, extensive experiments are also conducted on two fashion databases, Clothes-1M [18] and DeepFashion [12], for investigating the transfer learning capability.

3.1. Baseline Results

We report results on the validation set and private test set in Table 2, by using Inception-V1 [14], Inception-BN [8], Inception-V3 [15], and ResNet [5]. Experimental settings

Method	Validation			Private Test		
	R	P	F1	R	P	F1
Inception-BN	59.4	59.6	59.5	59.0	59.6	59.3
Inception-BN$*$	60.0	60.2	60.1	59.6	60.2	59.9
Inception-V1	59.9	60.1	60.0	59.5	60.1	59.8
Inception-V3	60.5	60.7	60.6	59.9	60.5	60.2
Resnet-101	59.7	59.9	59.8	59.3	59.9	59.5

Table 2. Baseline results on iFashion with precision, recall and F1. $*$ indicates ImageNet pre-trained.

Method	Training	Pre-trained	Accuracy
Inception-BN	50K clean	ImageNet	74.9
Inception-BN	50K clean	DeepFashion	76.4
Inception-BN	50K clean	Clothes-1M	77.5
Inception-BN	50K clean	iFashion	**78.9**
Inception-BN	1M + 50K	ImageNet	78.7
Inception-BN	1M + 50K	DeepFashion	78.3
Inception-BN	1M + 50K	iFashion	**80.5**
Xiao *et al*. [18]	1M + 50K	ImageNet	78.2
CleanNet [10]	1M + 50K	–	79.9
Patrini *et al*. [13]	1M + 50K	–	80.4

Table 3. Transfer learning on Clothes-50K.

are described in the supplementary material. From the Inception family, we found that a deeper network Inception-V3 outperforms Inception-V1 and Inception-BN on both validation and test sets. But surprisingly, the results of Inception-V1 are slightly better than that of Inception-BN. We hypothesize that such results may due to the complexity of our database which is significantly more difficult than single-label ones. Therefore, training CNNs from our dataset may require more local supervised information, and Inception-V1 and Inception-V3 have multiple loss functions to enhance local supervision. This challenge has not been fully investigated in the community, and may open new research interests on multi-label classification with hierarchical label structures. For data imbalance, we implemented a weighted binary cross entropy loss [12], and obtained 0.7% improvement with Resnet architecture.

3.2. Transfer Learning

To investigate the generalization ability of CNNs learned from iFashion, we compare it with various related databases used as source domain for transfer learning. We transfer the learned models from different source domains to new small-scale target domains: Clothes-50K [18] and DeepFashion [12]. Inception-BN [8] is used as the model, with regular training and fine-tuning schemes.

Transfer learning to Clothes-50K. Clothes-50K is a subset of Clothes-1M [18]. It has 50K training images from 14 categories and 14,312 validation images. Two groups of experiments are conducted. First, we train Inception-BN from 4 source databases, and fine-tune them on the Clothes-50K. Second, we train Inception-BN from the iFashion, DeepFashion and ImageNet, and then fine-tune the pre-trained

Method	Top1	Top3	Top5
WTBI [4]	–	43.7	66.3
DARN [7]	–	59.5	79.6
Yang *et al*. [19]	–	75.3	84.9
FashionNet [12]	–	82.6	90.2
Incpetion-BN (DeepFashion)	64.6	85.4	91.6
Incpetion-BN (Clothes-1M)	65.6	85.9	91.9
Incpetion-BN (ImageNet)	67.6	87.3	92.9
Incpetion-BN (iFashion)	69.2	**88.2**	**93.3**

Table 4. Transfer learning on DeepFashion.

CNNs by using the Clothes-1M and Clothes-50K sequentially, by following previous approaches implemented on the Clothes-1M and Clothes-50K. Results on the validation set of Clothes-50K are reported in Table 3.

As shown in Table 3, by using Clothes-50K as training data, the pre-trained model from iFashion obtains the best performance with 78.9% average accuracy. It outperforms the other three pre-trained models by large margins, particularly ImageNet with 74.9%. This suggests that with a similar data scale, our database has stronger generalization capability to fashion-related tasks than the object-centralized ImageNet. Compared with fashion-related DeepFashion or Clothes-1M, iFashion is larger in scale and has higher label quality, resulting in better generalization performance. More detailed comparisons and discussions are presented in the supplementary material.

In the second group of experiments, iFashion pre-trained models consistently outperform ImageNet and DeepFashion, but the impact of pre-trained models is decreased when the amount of training data is increased from 50K to 1M+. Furthermore, our result of 80.5% is better than those of recent approaches specifically designed to handle noisy data in Clothes-1M, while our model, empowered by iFashion, just employs a simple and straightforward fine-tuning method, with an off-the-shelf CNN.

Transfer learning on DeepFashion. DeepFashion [12] has 46 classes with 209,222 training images and 40,000 validation images, which were manually cleaned and annotated. We investigate the transfer capability of three *million-level* databases: ImageNet, Clothes-1M and iFashion. We train Inception-BN models individually on each of the three databases, and then fine-tune them on DeepFashion. Results are compared in Table 4.

The results are consistent with those on Clothes-50K: (i) all pre-trained models improved the performance over that of training from scratch; (ii) iFashion obtains the best performance on all terms, demonstrating its stronger capability for transfer learning; (iii) with iFashion pre-training, we can achieve state-of-the-art results on DeepFashion, by simply using an off-the-shelf Inception-BN. Interestingly, ImageNet pre-training has better performance than Clothes-1M, which may due to high-quality data with a number of overlapped fashion categories between ImageNet and Deepfashion, as analyzed in the supplementary material.

4. Conclusion

We present the iFashion dataset, which is the first known million-scale expertly curated image dataset with multi-label and fine-grained attributes. The aforementioned characteristics of the iFashion enable it to be relevant for real-world applications, particularly in fashion domain. The introduction of iFashion allows us to compare different approaches for multi-label learning, which we provide several baselines. Our experiments show that there is still large room to improve in this space. We also demonstrated the value of iFashion for transfer learning, where it outperforms the other well-known datasets on fashion recognition.

References

[1] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[4] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2016. CVPR.

[6] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[7] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015.

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[9] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *3DRR*, Sydney, Australia, 2013.

[10] K.-H. Lee, X. He, L. Zhang, and L. Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. *CoRR, abs/1711.07131*, 2017.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[12] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[13] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[17] X.-Z. Wu and Z.-H. Zhou. A unified view of multi-label performance measures. *CoRR, abs/1609.00288*, 2016.

[18] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

[19] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.