Leveraging Class Hierarchy in Fashion Classification

Hyunsoo Cho, Chaemin Ahn, Kang Min Yoo, Jinseok Seol, and Sang-goo Lee
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul, South Korea
{johyunsoo, chaae, kangminyoo, jamie, sglee}@europa.snu.ac.kr

Abstract

The online commerce market has been growing rapidly, spurring interest in the deep fashion domain from the research community. Among various tasks in the fashion domain, the classification problem is the vital one, because metadata extraction through fashion classification has tremendous industrial value. A flurry of recent deep-learning based models have been proposed for the task and have showed great performances but they fail to capture the hierarchical nature of fashion annotations, such as 'pant' and 'skirt' both having 'bottom' as the superordinate. In this preliminary work, we propose a novel fashion classification model that works in a hierarchical manner. Experimental results on large fashion datasets show that our intuition, taking into account hierarchical dependencies between class labels, can help improve performance.

1. Introduction

The online commerce market is growing exponentially. As of 2018, online apparel/fashion commerce occupies about 17% of the global online commerce market. Such rapid growth gave a dramatic push forward in the field of fashion image analysis. With the abundance of publicly available large-scale clothing datasets[14, 6, 7, 3], recent methods exploit the power of the deep neural network to achieve significant results in various fashion related tasks such as clothing item retrieval [7, 9], land mark detection [14, 21, 23], and classification [21, 19, 14]. Our prime interest is the classification problem in fashion domain, which is highly demanded in all stages of commerce. The ability to accurately extract metadata such as color, attributes and categories through classification can reduce the time, labor, and capital involved. Recent models in fashion classification show promising results but lack the consideration of the hierarchical nature of fashion annotations. Just as 'pants' and 'skirts' belong to 'bottoms', fashion items have a hierarchical structure with specific subsets and supersets. In this preliminary work, we propose a new classification model that reflects hierarchical relationships between fashion categories. This is based on the intuition that the model is more effective if they reflect hierarchical dependencies between class labels. To examine the effectiveness of our model, we conduct experiments on DeepFashion[14] dataset. Since DeepFashion dataset has no hierarchical annotation, we define subcategories by grouping some classes based on common features[1]. In this work, We propose a novel fashion classification model, which shows considerable performance improvements. Preliminary experiments hold our intuition that models are more effective when considering hierarchical dependencies between fashion labels.

2. Related Work

Fashion Image Classification

In recent years, clothing recognition has gained much attention due to its potential value. There are various tasks regarding clothing recognition such as category/attribute classification, landmark detection, clothes segmentation, and recommendation. Among various tasks above, the task of classifying fashion images is divided into several branches such as category, attribute, material, and style classification. A flurry of recent deep-learning based models have been proposed for the task and have showed great performances. FashionNet model[14] has improved the understanding of fashion item images with a large number of annotations by using multitask learning techniques. More recently, [13, 21] proposed a compact network for category classification and landmark detection using simple grammar topologies. By simultaneously learning classification and landmark detection, both models achieved remarkable model performance.

Hierarchical Classification

A hierarchical classification(HC) is a part of the classification task that maps input data into a defined hierarchical relationship, which can be represented in the form of a tree or a graph. To properly catch the relationship between layers, the HC algorithm must be able to label inputs to one or multiple paths in the class hierarchy. Early mod-

[1] https://bit.ly/2ZuT9Bf

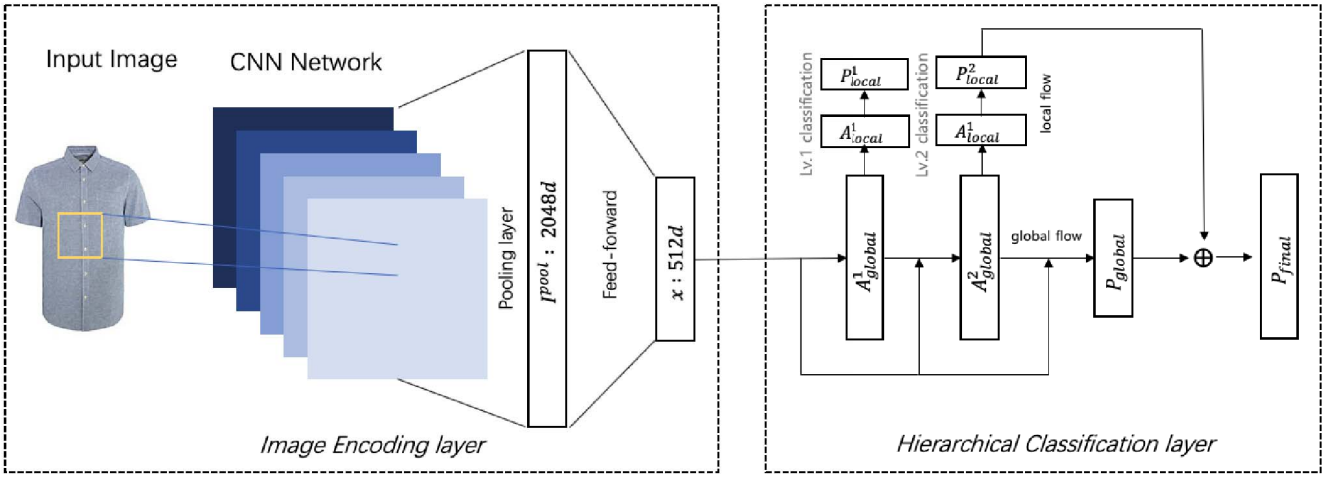

Figure 1. Overall Structure of our model

els, tried to solve this task in the local or global manner[18]. In the local method[12], the relationship between each hierarchy level is identified in distinct stages and later combined to generate the final classification. Global approaches for HC, on the other hand, usually consist of a single classifier capable of associating objects with their corresponding classes in the hierarchy as a whole [1, 2]. Structurally, global approaches are usually computationally cheaper and do not suffer from the error-propagation problem, but they are less likely to capture local information from the hierarchy. On the other hand, local approaches are much more suitable for extracting information from regions of the class hierarchy. Recently, instead of choosing a particular strategy [22] introduced Hierarchical Multi-label Classification Networks-Feedforward(HMCN-F), an approach that combines the merits of both local and global approaches and overcomes their shortcomings. As we will explain in subsequent sections, we choose HMCN-F as the basis for our model.

Multitask Learning

Multi-task learning(MTL) aims to improve learning efficiency and accuracy of multiple tasks in a single model as compared to training separate models for each task. MTL is based on the intuition that generalization can be improved by sharing domain information between complementary tasks, and have proven satisfactory performances in various fields, especially in computer vision. Various methods for MTL have been proposed[4, 16, 10, 17]. [10] uses a joint likelihood formulation to derive task weights based on the intrinsic uncertainty in each task. [4] proposes heuristics based on gradient magnitudes, and multi-agent reinforcement learning is used in [16]. Since MTL has proven its efficacy in various fields, we implement this powerful technique in our proposed model.

3. Model

Our approach has two components: *(1)* NN-based image encoder that maps an image into a fixed-sized embeddings, followed by a *(2)* hierarchical classifier over a set of annotations (categories). We implement baselines and our model on HMCN-F[22]. We transform the hierarchical multi-label classification problem as a type of multi-task learning problem and employ a MTL technique to optimize class weights. The overall structure of the proposed model is shown in Fig.

3.1. Image Encoding Layer

The input image is passed through a pretrained Convolutional Neural Network (CNN) to obtain a vector representation of size N×2048, where N is 7×7 grid image locations. Next, extracted vector goes through pooling layer($I^{pool} \in \mathbb{R}^{2048}$) and feed-forward network($x \in \mathbb{R}^{512}$), producing one reduced vector. In all cases, the CNN is pretrained and held fixed during the training.

3.2. Hierarchical Classification Layer

As informed earlier, the information in HMCN-F flows in two ways. *(1)* Global flow, which traverses all feed-forward networks from input to P_{global}. *(2)* Local flow, which expects the class of the hierarchy at each level from the global network at that level. Unlike the vanilla HMCN-F, only last-level local output and global output are used to calculate P_{final}. By reducing the dimension of the final prediction layer, it is not necessary to consider the violation loss in the original paper. Also, there is a slight improvement in performance as fewer dimensions are to be predicted.

3.2.1 Global and Local Flow

Global Flow sends and receives information to and from each level of the local flow. Given an input vector $x \in \mathbb{R}^{512}$, let A_G^1 denote the activations in the first level of the global flow layer. (C^h : set of categories of the h^{th} hierarchy level.)

$$A_G^1 = \phi(W_G^1 x + b_G^1) \in \mathbb{R}^s$$

where $W_G^1 \in \mathbb{R}^{s \times 512}, b_G^1 \in \mathbb{R}^s$ each indicates weight matrix and bias vector, and ϕ is a non-linear activation(e.g.,ReLU,tanh). Likewise, h^{th} global activations can be denoted as :

$$A_G^1 = \phi(W_G^h(A_G^{h-1} \odot x) + b_G^h)$$

where $\odot$ indicates the concatenation operator. The global prediction can be calculated by:

$$P_G = \sigma(W_G^{|H|+1} A_G^H + b_G^{|H|+1})$$

where $W_G^{|H|+1} \in \mathbb{R}^{|C^H| \times |A_G^{|H|}|}, b_G^{|H|+1} \in \mathbb{R}^{|C^h|}$ and σ refers to the sigmoid activation.(H indicates number of hierarchy level) Note that i^{th} elements of $P_G(P_G[i])$ can be interpreted as the probability of each class($P(C_i|x)$).

Local Flow can be calculated in the same manner as the global flow.

$$A_L^h = \phi(W_T^h x + b_T^h) \in \mathbb{R}^{|h|}$$

$$P_L^h = \sigma(W_L^h A_L^h + b_L^h)$$

where $W_T^h \in \mathbb{R}^{|A_L^h| \times |A_G^h|}, b_L \in \mathbb{R}^{|C^h|}$.

To fuse both local and global information into the final prediction, previously predicted logit values are summed over balance ratio β.(0.5 by default)

$$P_{final} = \beta P_L^{|H|} + (1 - \beta) P_G$$

3.2.2 Balanced Loss

In the original paper, arithmetic summation over local loss and global loss functions is minimized.

$$\mathcal{L}_L = \sum_{h=1}^{|H|} [\varepsilon(P_L^h, Y_L^h)]$$

$$\mathcal{L}_G = \varepsilon(P_G, Y_L^{|H|})$$

Where Y is the correct binary class vector for input x.

$$\mathcal{L}_{total} = \mathcal{L}_L + \mathcal{L}_G$$

However, simply adding loss function does not take into account the speed of training or the difficulty of each task. Therefore, we add a technique to find the optimum balance between losses using the method introduced in [10].

$$\mathcal{L}_{total} = \frac{1}{\sigma_1^2}\mathcal{L}_L + \frac{1}{\sigma_2^2}\mathcal{L}_G + log\sigma_1 + log\sigma_2$$

Methods	DeepFashion-test top-3	top-5	FLD
DARN[9]	59.48	79.58	×
Lu et al.[15]	86.72	92.51	×
Corbiere et al.[5]	86.30	92.80	×
effi+HMCN-F(5)	89.70	94.97	×
effi+HMCN-F(5)+MTL	90.17	95.09	×
effi+HMCN-F(10)	90.78	95.56	×
effi+HMCN-F(10)+MTL	**91.24**	**95.68**	×
FashionNet[14]	82.58	90.17	◯
Wang et al.[21]	90.99	95.78	◯
Li et al.[13]	**93.01**	**97.01**	◯

Table 1. Experiment Results on DeepFashion dataset.

4. Experiments

4.1. Experimental Settings

Recently, various large clothing datasets have been released publicly[14, 6, 7, 3]. Our experiments were conducted on the DeepFashion dataset due to its extensive coverage. All data is classified into 50 categories and consists of total of 289,222 images, of which 40,000 are test images.

Since DeepFashion dataset does not provide any hierarchical information, we additionally annotate the dataset by grouping 50 categories into certain number of parent categories. As fashion taxanomy can be ambiguous, we explore two different sets of category hierarchies. [2] We use Adam [11] optimizer with scheduled learning rate. All images were cut into regions of interest (ROI) and passed through pretrained EfficientNet-b5[20]. Each activation layer is followed by batch normalization, residual connections[8], and dropout of 70%. As for β in HMCN-F, we used 0.7. All hyperparameters have been determined through grid-based hyperparameter search. Experiments are run on 4x Tesla P100 GPUs.

4.2. Experiment Results

We compare our method with six different deep learning models in clothes recognition tasks. Table 1 summarizes the performance of different methods on category classification. Some of the models use Fashion Landmark Detection(FLD) to improve the model's performance. Our model can achieve considerable performance without FLD annotation. Empirically, we found that using MTL can improves performance and reduces convergence time. Also increasing the number of parent categories in datasets also benefits the model performance.

[2]Due to the page limit, a detailed explanation can be found at the link. https://bit.ly/2ZuT9Bf

5. Conclusion

In this preliminary work, we proposed a simple fashion category classification model that explores the hierarchical nature of fashion categories. We have shown that in the fashion area, performance improvements can be achieved by adding hierarchical information to datasets. Besides, by using MTL technique, our model can get better performance than the existing hierarchical classification model. As future work, we will experiment with various datasets and find a MTL technique that is better suited for hierarchical classification.

References

[1] Ricardo Cerri, Rodrigo C Barros, and Andre CPLF de Carvalho. A genetic algorithm for hierarchical multi-label classification. In *Proceedings of the 27th annual ACM symposium on applied computing*, pages 250–255. ACM, 2012.

[2] Ricardo Cerri, Rodrigo C Barros, André CPLF de Carvalho, and Alex A Freitas. A grammatical evolution algorithm for generation of hierarchical multi-label classification rules. In *2013 IEEE Congress on Evolutionary Computation*, pages 454–461. IEEE, 2013.

[3] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5315–5324, 2015.

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *arXiv preprint arXiv:1711.02257*, 2017.

[5] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2268–2274, 2017.

[6] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019.

[7] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

[10] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Svetlana Kiritchenko, Stan Matwin, and Fazel Famili. Hierarchical text categorization as a tool of associating genes with gene ontology codes. 2004.

[13] Peizhao Li, Yanjing Li, Xiaolong Jiang, and Xiantong Zhen. Two-stream multi-task network for fashion recognition. *arXiv preprint arXiv:1901.10172*, 2019.

[14] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[15] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5334–5343, 2017.

[16] Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017.

[17] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018.

[18] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[19] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2015.

[20] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[21] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018.

[22] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5225–5234, 2018.

[23] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 172–180. ACM, 2017.