Learning Image Representation via Attribute-Aware Attention Networks for Fashion Classification

Yongquan Wan[1,3(✉)], Cairong Yan[2], Bofeng Zhang[1], and Guobing Zou[1]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2] School of Computer Science and Technology, Donghua University, Shanghai, China
[3] School of Information Technology, Shanghai Jian Qiao University, Shanghai, China

Abstract. Attribute descriptions enrich the characteristics of fashion products, and they play an essential role in fashion image research. We propose a fashion classification model (M2Fashion) based on multi-modal data (text + image). It uses the intra-modal and inter-modal data correlation to locate relevant image regions under the guidance of attributes and the attention mechanism. Compared with traditional single-modal feature representation, learning embedding from multi-modal features can better reflect fine-grained image features. We adopt a multi-task learning framework that combines category classification and attribute prediction tasks. The extensive experimental result on the public dataset DeepFashion shows the superiority of our proposed M2Fashion compared with state-of-the-art methods. It achieves +1.3% top-3 accuracy rate improvement in the category classification task and +5.6%/+3.7% top-3 recall rate improvement in the attribute prediction of *part*/*shape*, respectively. A supplementary experiment on attribute-specific image retrieval on the DARN dataset also demonstrates the effectiveness of M2Fashion.

Keywordsss: Multi-modal · Classification · Prediction · Attention mechanism

1 Introduction

With the development of convolutional neural networks (CNN) and the publication of large-scale fashion datasets, significant progress has been made in fashion-related research, including fashion item recognition [1–3], fashion compatibility recommendation [4, 5], fashion attribute prediction [1, 6, 7] and fashion image retrieval [8–10]. Fashion category classification is a multi-class classification task, and fashion attribute prediction is a multi-label classification task. both of them generate helpful information for fashion items. Traditional classification methods usually only use the features learned from images as input and ignore the attribute information.

Figure 1 illustrates three fashion items in a fashion dataset. We can see that each image belongs to a category and has some attribute labels associated with it. The attribute labels of each fashion image reflect the category of the image. For example, an item with the '*strapless*' attribute is unlikely to be a pair of '*Jeans*', while an item with the '*mini*'

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13141, pp. 69–81, 2022.
https://doi.org/10.1007/978-3-030-98358-1_6

attribute is more likely to be a '*dress*'. In addition, there is a specific correlation between the attribute labels describing a fashion image, and they are not entirely independent. For example, '*denim*' and '*crochet*' will not be used to describe the same piece of clothing, while '*strapless*' and '*mini*' express the same piece of clothing because they are independent of each other. The use of labels and dependencies between labels helps to understand fashion items more accurately. Our goal is to use images and a group of known attribute labels to build a multi-modal classification model.

(a) category: dress (b) category: dress (c) category: Jeans

Fig. 1. Examples of fashion images and attributes. Image (a) and Image (b) share some attributes: *mini*, *strapless* and *sweetheart*, and they belong to the same category. Image (c) belongs to a separate category and has completely different attributes.

To support multi-modal interaction, we use two types of attention mechanisms to facilitate the interaction between visual and semantic information, i.e. an attribute-specific spatial attention module and an attribute-specific channel attention module. They enable the network to learn multi-modal features based on known attribute labels. In the model training phase, we represent the state of the labels as positive, negative or unknown to model them. Suppose we know the attribute state of image (a) and set it to true ('strapless') or false ('maxi'), the model can predict with a high degree of confidence that the image belongs to the 'upper body' category and has the 'mini' attribute. We compare our model with some competing methods on public datasets, which proves the model's superiority. The main contributions are as follows:

- We propose a fashion classification model (M2Fashion) based on multi-modal features. It is an attribute-guided attention-based model, which extracts more associated information between images and attributes to promote accurate fashion classification and attribute prediction. A channel attention module and a spatial attention module are integrated into the model for data fusion of two different modalities.
- We adopt a multi-task learning framework that combines category classification and attribute prediction tasks. Compared with other classification models, the attributes in our model are not independent, and their relationship is contained in the attribute hierarchy.
- Extensive experiments are carried out to compare the proposed model with several state-of-the-art models on public datasets. Experimental results show the superiority of the proposed model. In addition, M2Fashion is applied to an attribute-specific image retrieval tasks by removing the final classifier. This supplementary experiment also demonstrates the effectiveness of our model.

2 Related Work

Attribute Learning. The existing attribute learning methods can be categorized into two groups: 1) visual feature-based [9, 10]. They embed images in a common low-dimensional space and use the feature vectors in the low-dimensional space for attribute classification. 2) visual-semantic feature-based [11–13]. They learn joint representation by exploring the correlation between multi-modal content. Some of these methods use semantic information from attributes or annotated text to extract saliency or visual attention from the image. The above studies all learn visual/semantic features but ignore the relationship between attributes. Our work aims to mine the inner correlation of multiple attributes to learn fine-grained image representations.

Attention-Based Models. In recent years, the attention mechanism is widely used in computer vision and natural language processing. This technology has also been researched and applied in the field of fashion. Ji et al. [14] proposed a tag-based attention mechanism and a context-based attention mechanism to improve the performance of cross-domain retrieval of fashion images. Li et al. [15] proposed a joint attribute detection and visual attention framework for clothes image captioning. Ma et al. [16] proposed an attribute feature embedding network, which learns attribute-based embedding in an end-to-end manner to measure the attribute-specified fine-grained similarity of fashion items. Inspired by the success of the attention mechanism, we proposed to use two attribute-aware attention modules for fine-grained image classification tasks.

Multi-task Learning. Since it was proposed, multi-task learning (MTL) has achieved many successes in several domains, such as image classification with landmark detection [17], attribute-enhanced recipe retrieval [18], and visual question answering [19]. To explore the intrinsic correlation of attributes to obtain more reliable prediction results, we are motivated to build a multi-task framework to model the correlation and common representation of categories and multiple attributes of fashion images.

3 Methodology

3.1 Problem Formulation

Given a set of fashion items denoted by $D = \{(x_1, A_1), ..., (x_n, A_n)\}$, where $x_i (1 \leq i \leq n)$ is the i-th image, $x_i \in \mathbb{R}^{c \times h \times w}$ (c, h, and w are the number of channel, height, and weight respectively), $A_i = [a_{i1}, a_{i2}, ..., a_{iK}]$ is a multi-hot attribute vector which describes the image appearance with K semantic attributes, $a_{ij} \in \{-1, 0, 1\}$ $(1 \leq j \leq \mathrm{K})$, and K is the number of all attributes. The attribute set is denoted as $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_K\}$. The goal of our model is to map the unimodal representation from images and attributes to a joint semantic space, and learn a classifier $f(\bullet)$ in the joint space so that $y = f(x, A; \theta)$. In category classification tasks, y denotes the predicted image category, and in attribute prediction tasks, y denotes predicted attribute labels.

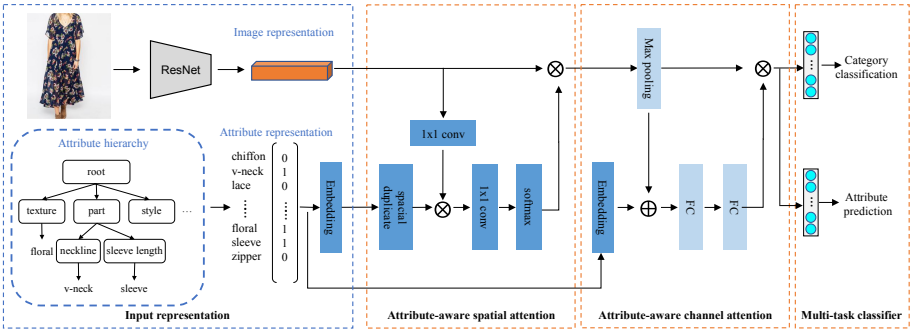

Fig. 2. The framework of our proposed model. It is made up of four key components, input representation module, attribute-aware spatial attention module, attribute-aware channel attention module, and multi-task classifier.

3.2 Network Structure

Figure 2 illustrates the framework of our model. It consists of four key components: input representation module, attribute-aware spatial attention (ASA) module, attribute-aware channel attention (ACA) module, and multi-task classifier. For an input image, the image embedding vector is extracted using ResNet pre-trained on ImageNet. Then, to learn the fine-grained features of the image, we use the image and multiple attributes to learn the feature representation. We adopt the architecture of [16] but add some changes to the method. In their work, spatial attention and channel attention of images guided by attributes are generated by embedding one attribute category such as 'sleeve_length'. In contrast, our model combines images and attribute values such as '3/4 sleeve' to generate attribute-guided attention. Our intuition is that images with the same attribute values will have more similar features. After that, these new attribute-aware features are fed into the attribute classifier.

Input Representation. To represent an image, we use ResNet, a CNN model pre-trained on ImageNet, as the backbone network. To maintain the spatial information of the image, we remove the last fully connected (FC) layer in CNN. Given an image $x_i \in \mathbb{R}^{h\times w\times 3}$, the feature extractor outputs a vector $\overline{x}_i \in \mathbb{R}^{h\times w\times d}$, where $h \times w$ is the size of the feature graph, and d is the number of channels. We represent the attribute label status as positive, negative, or unknown. They are represented by 1, -1 and 0, respectively. For an image x_i, we collect a set of labels embedded in A_i, the j-th element in A_i means the i-th image has the attribute a_{ij}. Attribute label embeddings $\overline{A}_i$ are learned from an embedding layer of size $d \times$ K.

$$\overline{A}_i = f_1(A_i) = \delta(W_{a1}A_i), \tag{1}$$

where $W_{a1} \in \mathbb{R}^{d\times K}$ denotes transformation matrix, δ denotes the tanh activation function. Note that we broadcast $\overline{A}_i$ along the height and width dimension so that its shape is compatible to the image feature map $\overline{x}_i$.

Attribute-Aware Spatial Attention. An attribute is related to a specific visual region of the fashion image. For example, the attribute '*3/4 sleeve*' usually appears on either side of the middle area in the image, and to learn attribute-specific features such as '*sleeve length*', the regions around the sleeve will receive more attention. To calculate the attribute-specific image space attention, instead of using a single attribute category to guide attention, we use multiple attribute values to generate attribute embedding. These values are organized into a hierarchical structure, called attribute hierarchy.

Specifically, for an image x_i and its attribute labels A_i, we use I and T_1 to represent $\overline{x}_i$ and $\overline{A}_i$, respectively. First, we get the attribute guided spatial attention feature vector denoted as V_s, obtained by calculating the weighted average of the input image features according to the attribute label embedding. For image embedding I, we employ a convolution layer with d 1×1 convolutional kernels following a nonlinear tanh activation function to transform the dimension of the image to d. The mapped image feature vector is expressed as

$$f_2(I) = \delta(W_{v1}I), \tag{2}$$

where W_{v1} denotes a convolutional layer containing d 1×1 convolution kernels, and δ denotes the tanh activation function.

The attended image feature vector is fused with attribute feature using element-wise product followed with an activation function.

$$f_s(I, T_1) = \delta(W_{v2}(f_2(I) \odot T), \tag{3}$$

where $\odot$ denotes element-wise product operation, W_{v2} is 1×1 convolutional layer, and δ denotes the tanh activation function. The attention weight is obtained through the softmax activation function.

$$\alpha_l^s = \frac{\exp(f_s(I_l, T_1))}{\sum_j^{h \times w} \exp(f_s(I_j, T_1))}. \tag{4}$$

Then, the spatial attention feature vector under the attention of attribute A_i can be obtained by the following calculation.

$$V_s = \sum\nolimits_l^{h \times w} \alpha_l^s I_l, \tag{5}$$

where $\alpha_l^s \in R^{h \times w}$ is the attention weight, and I_l is the image feature at location l.

Attribute-Aware Channel Attention. We adopt the attention mechanism of Ma et al. [16] with one modification. In their work, they apply sum pooling on the output from ASEN module. In contrast, we adopt global max pooling on the feature map V_s to concentrate only on discriminative areas. For the attribute A_i, we employ a separate attribute embedding layer to generate an embedding vector with the same dimension as V_s,

$$\tilde{A}_i = f_3(A_i) = \delta(W_{a2}A_i), \tag{6}$$

where $w_{a2} \in \mathbb{R}^{c \times n}$ is the embedding parameter, and δ is the tanh activation function. For the convenience of understanding, we use T_2 to represent $\widetilde{\mathrm{A}}_i$. The spatial attended features and attribute embedding are fused by concatenation, then fed into two sequential FC layers to generate the attribute-aware channel attended feature. The attention weight $\alpha^c \in \mathbb{R}^c$ is calculated by

$$\alpha^c = \sigma(W_{c2}\sigma(W_{c1}[T_2, V_s])), \tag{7}$$

Where [,] represents the concatenation operation, σ represents the sigmoid activation function, and W_{c1} and W_{c2} are parameters of the FC layer. For simplicity of understanding, the bias in the formula is removed. The final output of ACA is obtained by the element-wise product of I_s and attention weight α_c.

$$V_c = \alpha^c \odot V_s. \tag{8}$$

Finally, we further employ an FC layer over V_c to generate the attribute-guided feature of the given image with known image labels.

$$Z = WV_c + b, \tag{9}$$

where $W \in \mathbb{R}^{c \times c}$ is the transformation matrix, and $b \in \mathbb{R}^c$ is the bias.

Multi-task Learning. In this paper, the MTL framework is used to predict the categories and attributes of images. We share feature vectors in two tasks, category classification and attribute prediction, which helps to share knowledge and distinguish subtle differences between different tasks. At the end of the network, we add two different branches, one for predicting categories of images and the other for predicting attributes of images. The shared attribute-guided image features output is fed to two branches, respectively. We use the cross-entropy loss for category classification, denoted as

$$L_{category} = -\frac{1}{N}\sum_{i=1}^{N}\{y_i^c \log(P(\hat{y}_i^c|Z_i)) + (1 - y_i)\log(1 - P(\hat{y}_i^c|Z_i))\}. \tag{10}$$

The output of the attribute prediction branch is passed into a sigmoid layer to squeeze the output between [0,1] and output $\hat{a}_j$. We use the binary cross-entropy loss for attribute prediction, denoted as

$$L_{attribute} = \sum_{j=1}^{K} a_j \log(p(\hat{a}_j|x_i)) + (1 - a_j)\log(1 - p(\hat{a}_j|x_i)), \tag{11}$$

where a_j is the j-th ground truth of the binary attribute label, $p(\hat{a}_j|x_i)$ is a component of $Y = [y_1, \cdots, y_k]$, and Y is the predicted attribute distribution.

3.3 Label Mask Training

We adopt the strategy of label masking training proposed in [20] to learn the correlation between labels and allow the model to perform multiple label classification with given

partial labels. In the process of training, we mask a certain number of labels randomly and use the ground truth of other labels to predict masked labels. For K possible labels, we set certain labels Y_u as unknown labels for a particular sample, where $|Y_u|$ is a random number between 0.25K and K. Y_u are randomly sampled from all available labels Y, and their state is set as unknown. The remaining labels are known and denoted as Y_k. These labels in the known state will be used as input to the model along with the image, and our model predicts labels in the unknown state. In the training process, some labels are randomly masked as unknown, and the model learn the combination association of different known status labels. After the label mask training is incorporated, Eq. (11) is modified as

$$L_{attribute} = \sum_{j=1}^{K} E\{CE(Y_u, \hat{Y}_u | Y_k)\}. \tag{12}$$

3.4 Triplet Network Training

We use the triplet network shown in Fig. 3 to train our model, aiming to learn effective embedding and similarity measurements to minimize the distance between anchor and positive samples and maximize the distance between anchor and negative ones.

The construction process of the input data in the triplet network is as follows. Given an image triplet $\{x^a, x^p, x^n\}$, x^a is the anchor image, x^p is the positive image, and x^n is the negative image. The positive example image has at least one attribute that is the same as the anchor image, while the negative example does not have any attribute the same as the anchor image. Let $\{Z^a, Z^p, Z^n\}$ be the attribute attended feature embedding triplet. The similarity is defined as cosine similarity.

$$sim(Z^a, Z^p) = \frac{Z^a \cdot Z^p}{\|Z^a\| + \|Z^p\|}, sim(Z^a, Z^n) = \frac{Z^a \cdot Z^n}{\|Z^a\| + \|Z^n\|}. \tag{13}$$

We force the similarity between the anchor and the positive samples to be greater than the similarity between the anchor and the negative samples, i.e., $sim(Z^a, Z^p) > sim(Z^a, Z^n)$. Then we define a triplet ranking loss function based on hinge loss as

$$L_{tri} = max\{0, sim(Z^a, Z^p) - sim(Z^a, Z^n) + m\}, \tag{14}$$

where m represents the margin between two similarities. The total loss is defined as

$$L_{total} = L_{category} + \lambda L_{attribute} + \gamma L_{tri}, \tag{15}$$

where λ and γ are parameters that balance the contribution of all losses.

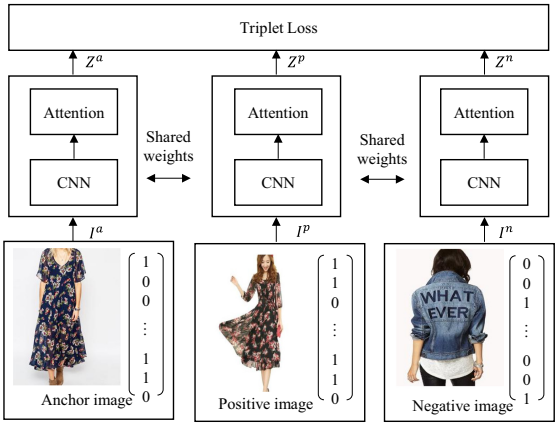

Fig. 3. The triplet network structure used to train our model.

4 Experiments

4.1 Experiments Settings

Datasets. We conduct our experiments on a public dataset Deepfashion [1], a large-scale clothes dataset. We choose its Category and Attribute Prediction benchmark (abbreviated as DeepFashion-C) that is more suitable for our tasks. DeepFashion-C contains 289,222 clothes images in 46 categories and five attribute categories with 1,000 attribute values. Each image is annotated with only one category and several attributes. We adopt the same train-valid-test division as [1].

Metrics. For image category classification, top-k accuracy is usually adopted as the evaluation metric. For image attribute prediction, the top-k recall rate used in [1] is traditionally used as the evaluation metric.

Implementation Details. The proposed model is implemented in the Pytorch framework with an NVIDIA GeForce GTX 1080Ti GPU. We use the ResNet 50 network pre-trained on ImageNet for feature extraction. The images are resized to 224×224. We use a 1×1 convolutional layer to reduce the dimension of the feature vector to 512. The multi-hot vector of the attributes is transformed to 512-dimensional vectors by an embedding layer followed by the tanh activation function. Then the image and attribute features are used to obtain spatial attention through the dot product operation. In the ACA module, we use a separated attribute embedding layer. We use SGD to train the triplet network, the total epoch is set to 20. The learning rate is 1e-5 and decays at the rate of 0.95 every epoch. We empirically set α to 1 and γ to 0.5 in Eq. (15).

Baselines. We conduct comparative tests with some baseline models. All models use the same triple sampling method for fine-tuning, but the training methods are different. WTBI [21] first trains a generalized similarity model, and then fine-tunes each type of clothing to obtain a class-independent model. DARN [8] constructs a tree structure for all attributes to form a semantic representation space of clothing images. FashionNet [1]

extracts features and landmark location information from images, and combines them for training to predict image categories and attributes. Corbiere [9] uses weak label information and images crawled from the Internet to make dot products and predicts the probability of each word in the vocabulary. Attentive [3] uses a two-way convolutional recursive neural network to improve classification through landmark-aware attention and category-driven attention. Upsampling [22] increases the resolution of the feature map through up-sampling and uses the predicted landmark location as a reference to improve classification.

4.2 Experiment Results

We validate the performance of our model on the DeepFashion-C dataset, and Table 1 summarizes the performance of different methods in terms of top-k ($k = 3, 5$) recall rate for fashion classification and attribute prediction. Some clothing classification and attribute recognition results are show in Fig. 4. The following observations can be obtained.

Table 1. Performance comparison of different models on DeepFashion-C dataset

Models	Category		Texture		Fabric		Shape		Part		Style		All	
	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5
WTBI	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet	82.58	90.17	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02	66.43	73.16	45.52	54.61
Corbiere	86.30	92.80	53.60	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
Attentive	90.99	95.78	50.31	65.48	40.31	48.23	53.32	61.05	40.65	56.32	68.70	**74.25**	51.53	60.95
Upsampling	91.16	96.12	56.17	65.83	**43.20**	53.52	58.28	67.80	46.97	57.43	**68.82**	74.13	54.69	63.74
Ours w/o ASA	91.89	96.13	55.61	65.17	40.12	54.73	59.70	68.98	49.17	59.19	64.48	71.82	52.64	62.37
Ours w/o ACA	90.12	95.15	54.73	64.82	39.45	53.27	58.21	66.72	47.87	57.53	60.54	68.61	50.47	60.07
Ours	**92.33**	**96.65**	**56.89**	**66.31**	40.42	**55.83**	**60.42**	**69.87**	**49.62**	**61.17**	65.38	72.79	**54.85**	**64.76**

- Our model outperforms all competitors in the category classification task and the attribute prediction task. For the former, our model improves the top-3 accuracy rate by 1.3%. For the latter, our model also improves the recall rate.
- We evaluate our model using only one attention module and get two variants: M2Fashion w/o ASA and M2Fashion w/o ACA. The former employs global max pooling instead of an attribute-aware spatial attention model to generate features. The latter utilizes vector V_s as the attribute-guide feature vector directly. We can see that removing the ASA or ACA module reduces the performance of the two subtasks, showing the effectiveness of both ASA and ACA modules.
- The classification task has a more significant impact on the *part*-related attribute prediction (+5.6% of top-3 recall rate) and the *shape*-related attribute prediction (+3.7%

Fig. 4. Results of clothing category classification and attribute prediction on DeepFashion-C dataset. The correct predictions are marked in green, and the wrong predictions are marked in red.

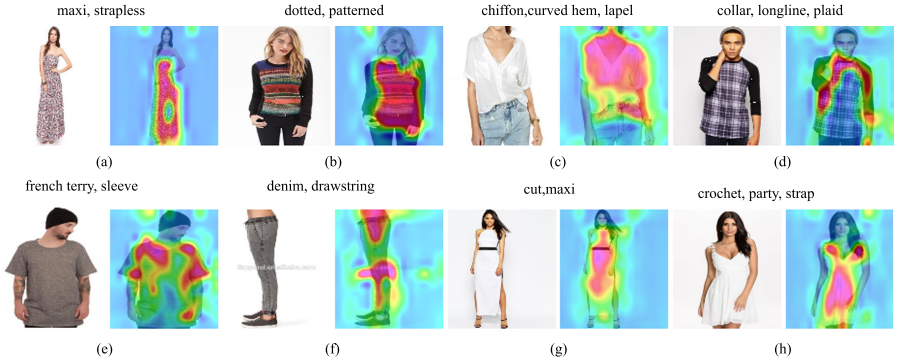

Fig. 5. Visualization of the attribute-aware spatial attention on DeepFashion-C.

top-3 recall rate) than the *texture*-related attribute prediction (+1.3% top-3 recall rate). It does not perform well on the *style*-related attribute prediction and the *fabric*-related attributes prediction because it is hard to focus the attention on these two attributes on the images. The classification of clothing is more dependent on the shape characteristics of clothing, and clothing classification can also promote the understanding of *shape*-related attributes.

4.3 Attention Visualization

Visualization of our attention mechanisms can be found in Fig. 5. We observe that the learned attention gives a higher response in the attribute-related areas, which shows that the attention helps find out which areas are relative to the given attribute. According to our observations, the attributes related to '*part*', such as '*maxi*' in Fig. 5(a) and '*sleeve*' in Fig. 5(e), are more likely to highlight local visual features. The attention map of attributes related to '*material*' or '*style*' focuses on the entire clothing.

4.4 Impact of Joint Learning and the Pooling Methods

The Impact of Joint Learning. We explore the correlation between category classification and attribute prediction. As shown in Table 2 (top), the results show that the joint learning of categories and attributes improves the accuracy of the two tasks. We found that after adopting the multi-task learning framework, in the classification task, the top-3 accuracy is increased by 4.1%, and the top-5 accuracy is increased by 3.0%; in the attribute prediction task, the top-3 recall rate is increased by 11.7%, and the top-5 recall rate is increased by 12.2%.

The Impact of Global Max Pooling. We use global max pooling instead of global average pooling to capture global context information. Global max pooling is sensitive to discriminative local features. The function of global maximum pooling is verified by ablation experiments. The results are shown in Table 2 (bottom). Global max pooling improves the recall rate of category classification and attribute prediction.

Table 2. Performance comparison of different learning methods and pooling methods.

Methods	Category		All	
	Top-3	Top-5	Top-3	Top-5
Category only	88.64	93.84	–	–
Attribute only	–	–	48.65	57.74
Category + attribute	**92.23**	**96.65**	**54.35**	**64.76**
GAP	91.72	95.84	53.63	62.84
GMP	**92.23**	**96.65**	**54.35**	**64.76**

Table 3. Performance comparison of attribute-specific fashion retrieval on DARN using MAP

Models	Category	Clothes button	Clothes color	Clothes length	Clothes pattern	Clothes shape	Collar shape	Sleeve length	Sleeve shape	All
Triplet	23.59	38.07	16.83	39.77	49.56	47.00	23.43	68.49	56.48	40.14
CSN	34.10	44.32	48.38	53.68	54.09	56.32	31.82	78.05	58.76	50.86
ASEN	36.69	46.96	**51.35**	56.47	54.49	60.02	34.18	80.11	60.04	53.31
Ours	**36.91**	**48.03**	51.14	**57.51**	**56.09**	**60.77**	**35.05**	**81.13**	**62.23**	**54.29**

4.5 A Case: Attribute-Specific Image Retrieval

The learned model can be applied to attribute-specific image retrieval tasks by removing the final classifier. For instance, given a query with an image and two labels *v-neckline* and *floral*, the model returns top-k similar images with these two labels.

We conduct the experiment on DARN [8] dataset, which contains about 253983 upper-clothing images and has a total of 9 attributes and 179 attribute values. We randomly divided the dataset into 8:1:1 for training, validation and test. Similar to [16], we use the metric of mean average precision (MAP) for evaluation. Following baselines are considered for comparison: Triplet Network, Conditional Similarity Network (CSN) [23], and Attribute specific embedding network (ASEN) [16].

Table 3 shows the results of attribute-specific image retrieval tasks on the DARN dataset. We can see that (1) the triplet network that learns the universal embedding space performs the worst; (2) our proposed M2Fashion outperforms other baselines. We attribute the better performance to the fact that M2Fashion uses multiple attribute labels and label masks to learn the association between labels and the attention of labels with images. In contrast, ASEN uses a single attribute category to guide attention.

5 Conclusions

In this paper, we explore fine-grained fashion image embedding to capture multi-modal content for fashion categorization. The proposed model adopts the visual-text attention mechanism to capture the association between different modal data and effectively uses any number of partial labels to perform multi-label and multi-class classification tasks. It also helps to discover how different attributes focus on specific areas of an image. In the future, we will study the impact of the hierarchical structure of attributes on the model and extend the model to hierarchical attribute prediction.

References

1. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: CVPR, pp. 1096–1104 (2016)
2. Dong, Q., Gong, S., Zhu, X.: Multi-task curriculum transfer deep learning of clothing attributes. In: WACV, pp. 520–529 (2017)
3. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: CVPR, pp. 4271–4280 (2018)
4. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional LSTMs. In: MM, pp. 1078–1086 (2017)
5. Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V.W., Liu, Q.: Explainable fashion recommendation: a semantic attribute region guided approach. In: IJCAI, pp. 4681–4688 (2019)
6. Zhang, S., Song, Z., Cao, X., Zhang, H., Zhou, J.: Task-aware attention model for clothing attribute prediction. IEEE Trans. Circ. Syst. Video Technol. (TCSVT) **30**(4), 1051–1064 (2020)
7. Chen, M., Qin, Y., Qi, L., Sun, Y.: Improving fashion landmark detection by dual attention feature enhancement. In: ICCV Workshop, pp. 3101–3104 (2019)
8. Huang, J., Feris, R., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: ICCV, pp. 1062–1070 (2015)
9. Corbiere, C., Ben-Younes, H., Rame, A., Ollion, C.: Leveraging weakly annotated data for fashion image retrieval and label prediction. In: ICCV Workshop, pp. 2268–2274 (2017)
10. Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: CVPR, pp. 5315–5324 (2015)

11. Liao, L., He, X., Zhao, B., Ngo, C.W., Chua, T.S.: Interpretable multimodal retrieval for fashion products. In: MM, pp. 1571–1579 (2018)
12. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y.: Automatic spatially-aware fashion concept discovery. In: ICCV, pp. 1463–1471 (2017)
13. Ferreira, B.Q., Costeira, J.P., Sousa, R.G., Gui, L.Y., Gomes, J.P.: Pose guided attention for multi-label fashion image classification. In: ICCV Workshop, pp. 3125–3128 (2019)
14. Ji, X., Wang, W., Zhang, M., Yang, Y.: Cross-domain image retrieval with attention modeling. In: MM, pp. 1654–1662 (2017)
15. Li, X., Ye, Z., Zhang, Z., Zhao, M.: Clothes image caption generation with attribute detection and visual attention model. Pattern Recogn. Lett. **141**, 68–74 (2021)
16. Ma, Z., Dong, J., Long, Z., Zhang, Y., He, Y., Xue, H.: Fine-grained fashion similarity learning by attribute-specific embedding network. In: AAAI, pp. 11741–11748 (2020)
17. Li, P., Li, Y., Jiang, X., Zhen, X.: Two-stream multi-task network for fashion recognition. In: ICIP, pp. 3038–3042 (2019)
18. Min, W., Jiang, S., Sang, J.: Being a supercook: joint food attributes and multimodal content modeling for recipe retrieval and exploration. IEEE Trans. Multimedia **19**(5), 1100–1113 (2017)
19. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: CVPR, pp. 10437–10446 (2020)
20. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: CVPR, pp. 16478–16488 (2021)
21. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: matching street clothing photos in online shops. In: ICCV, pp. 3343–3351 (2015)
22. Liu, J., Lu, H.: Deep fashion analysis with feature map upsampling and landmark-driven attention. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11131, pp. 30–36. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11015-4_4
23. Veit, A., Belongie, S., Karaletsos, T.: Conditional similarity networks. In: CVPR (2017)