Personalized Fashion Recommendation using Pairwise Attention

Donnaphat Trakulwaranont[1,2][0000−0002−3011−555X], Marc A. Kastner[2][0000−0002−9193−5973], and Shin'ichi Satoh[2,1][0000−0001−6995−6447]

[1] The University of Tokyo, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
{eiam,mkastner,satoh}@nii.ac.jp

Abstract. The e-commerce fashion industry is booming and comes with the need for proper search and recommendation. However, sufficient user personalization is still a challenging task. In this paper, we introduce a personalized fashion recommendation system based on high-dimensional input of user- and environment information. The proposed framework is used to estimate suitable categories and style of clothing depending on customized settings such as body type, age, occasion, or season. The goal is to recommend a full fitting outfit from the estimated suggestions. However, various personal attributes add up to a high dimensionality, and datasets are often very unbalanced or biased, making it difficult to do a proper recommendation. To solve this, we propose a pairwise-attention module to improve the performance of our framework. Our model can improve the performance up to 53.29% over the comparison method on MSE, mAP, and Recall. Moreover, in a subjective evaluation with human participants, the recommendations of the proposed method are preferred over the comparison method.

Keywords: Recommendation Systems · Personalized Recommendation · Fashion Media

1 Introduction

Clothing is one of the first impressions that people get from one another. Fashion tells a story about what we are and what we want to be. This is reflected in the fashion industry, which reported a growth of revenue from 1.3 trillion U.S. dollars (in 2012) to 1.8 trillion U.S. dollars (in 2019)[3]. More recently, fashion trends were further influenced by digital disruption and cross-border challenges. A major fashion trend is to get more personalized [3], resulting in online fashion retailers to invest and use more recent machine learning-based technology. Further, it is becoming more global and digital [2], accelerated through the global pandemic [1]. However, online fashion shopping leads to obstacles, such as the difficulty of judging whether a certain piece of clothing looks good on oneself or which kind of clothing item is more suitable.

[3] https://www.oberlo.com/statistics/apparel-industry-statistics

To help customers to make a decision in their fashion shopping, traditional methods use the purchase history and examples of clothing items to make better suggestions. However, this cannot suit clothing recommendations to specific users or scenarios, especially for rare occasions or lesser-known users. Some work [9,10] started to introduce body measurements or 3D body shape as an input to suggest suitable clothing. Others [21] proposed to use more personal information such as event and gender information, and also use preferred outfit images to query the outfit from the database based on similarity. Although these works can achieve some of their objectives, there still are some limitations and drawbacks such as missing important personal information, a low variety of clothing types, and limitations in suggested clothing category and style.

To conquer these limitations, our objective is to improve personalization in fashion recommendation systems by including a high number of personal attributes (such as age, ethnicity, and body shape) as well as environmental information (such as occasion or season) at the same time. Media datasets are commonly unbalanced and often contain only sufficient data for a part of users (e.g., a common bias is towards white males, young age groups, and so on). Due to this, a high dimensionality of user inputs becomes an issue. To solve this, we propose a pairwise attention module to combine each attribute and improve the training performance for lesser-known combinations of queries. With this, we can receive a more personalized recommendation system for suggesting types and styles of clothing. The proposed method is evaluated and compared to a comparison method [10], showing promising performance. We further employed a subjective evaluation with a user study, where a majority of participants preferred the recommendations of our system compared to the comparison method.

The main contribution of this paper can be summarized as follows:

- We propose a multi-attribute recommendation-query framework to suggest the outfit most appropriate to a specific person on a specific occasion/season.
- To solve issues with data imbalance for lesser-known input combinations, we propose a novel pairwise attention module, which is able used to better understand the connection of existing data samples.
- We evaluate the framework both in comparison with an existing method on quantitative measures, as well as a subjective evaluation with human participants.

2 Related work

In the following, we discuss existing related work on recommendation systems, both general-purpose as well as those targeting fashion media.

General-purpose recommendation. Zheng et al. [22] introduce a recommendation system called DeepCoNN, which works with text-based user reviews. They used a pre-trained word embedding to embed text information for input to CNN layers, to extract multiple levels of features from text input. Finally, they use

a Factorization Machine [17] (FM) as a rating estimator. Rawat et al. [16] propose ConTagNet for recommending tags based on input image and user-context using the YFCC100M [20] dataset. They use AlexNet [12] to extract features from input images and a custom neural network to extract features from tag information. After that, they concatenate both features to perform a multi-label classification to predict tag scores. He et al. [8] use a CNN to learn the interaction between user and item information using Yelp reviews and Gowalla check-in data. They embed user and item information to each feature vector, and generate an interaction map using outer product operation. Then, using ConvNCF which is a stack of six convolutional layers to learn the correlations between user and item on interaction map, they predict the item recommendation.

Personalized fashion recommendation. Compared to the general-purpose recommendation, personalized fashion recommendation is usually more closely tight to user information as not only age, and body type, but also the target occasion play a very crucial role in deciding the right outfit. "What dress fit me best" [9] proposes fashion item recommendation based on a correlation between body shape and clothing style. They construct a celebrity dataset "Style4BodyShape", featuring body measurement, stylist information, and related fashion outfits. They propose a method that calculates body shape into seven types and do a personalized style suggestion on top of that. Hsiao et al. [10] propose ViBE, recommending clothing based on the relation between body shape, clothing, and clothing attribute. However, instead of using only body measurement, they also include 3D body shape images in the proposed method to predict more close recommendations. Most recently, Fashionist [21] do personalization by including more user information such as gender and occasion. They also use a user's preference based on the preferred outfit image, then use the visual preference modeling to extract the semantic information from the preferred outfit image.

In this work, we are inspired by this variety of work introducing additional attributes into the personalized fashion recommendation. However, we also note dataset imbalance and insufficient data for lesser observed input combinations, especially if introducing a high number of customizable personal attributes. Thus, the target of this research is to solve these remaining issues and propose a more robust personalized recommendation system.

3 Proposed framework

Our goal is to create a recommendation system to suggest an outfit based on personal- and environmental information. Our proposed method can be divided into three stages: First, we augment the data of existing fashion datasets by extraction. Second, in the recommendation stage, clothing categories and attributes are suggested based on the high-dimensional input of personal attributes, wearing occasion, and wearing season. Third, in the query stage, an adequate outfit that matches the recommended output from the recommendation stage is selected from a large outfit dataset. The overall structure is shown in Fig. 1.

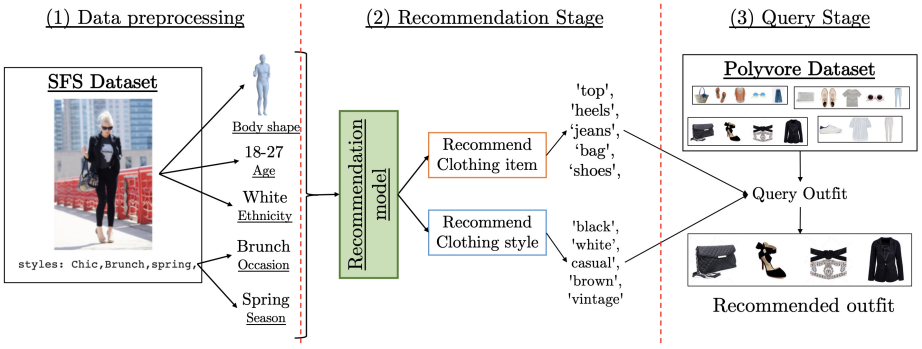

Fig. 1. Overview of proposed framework. The proposed framework consists of three stages: (1) Dataset preprocessing, extracting visual information from existing fashion datasets to augment the usable data, (2) Recommendation stage, that uses a recommendation model to predict the clothing item and attribute, (3) Query stage that uses output from recommendation stage to query outfit as overall system output.

3.1 Dataset preprocessing

Browsing existing fashion datasets [5,6] quickly reveals some limitations for personalized recommendation. There often is a dataset imbalance, with some occasions or gender/age combinations being highly available while there are almost no samples for other combinations. Further, by its nature fashion attributes are long-tailed, making them hard to train and recommend. This makes the data noisy, often also resulting in inconsistent or incomplete annotations.

To solve these limitations, we extract additional data from all available images to augment the existing dataset. Using existing methods [19,18], we estimate personal attributes from user images. We generate estimates for ethnicity, age, and body shape type.

3.2 Recommendation stage

The architecture of the recommendation stage is designed with a stack of convolution layers followed by a BatchNorm layer, a ReLU layer, and Residual layers [7] to form feature extractor Fe. To deal with the high number of inputs, we form a feature extractor for each type of input and then combine it after feature extraction as shown in the left part of Fig. 2.

For the pairwise attention module, the main objective is to generate a weight attention score for each type of input, such as different user- or environment information. Therefore, when the model encounters each combination of input, it can properly weigh each input feature, being able to give better recommendations for lesser-known combinations. A weight attention score is generated using multiplication between feature vectors and using the Sigmoid function to map the value to 0 to 1. The structure of the pairwise attention module is shown in

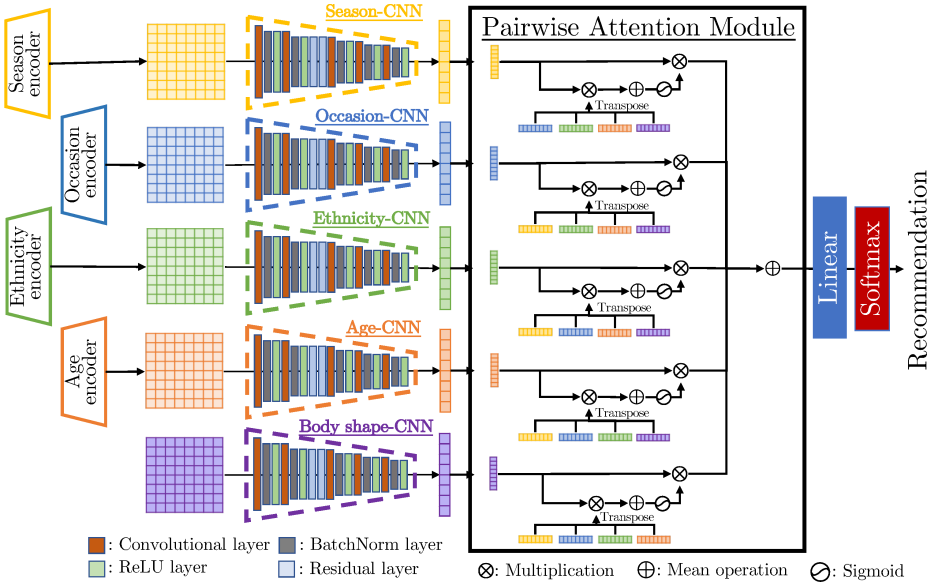

Fig. 2. Recommendation model architecture which consists of two part: (1) feature extraction part that uses convolution layers with BatchNorm, ReLU and Residual layers to extract features, and (2) feature combination part is a pairwise attention module which is used for generate weight attention score for each type of input data.

the right-side box in Fig. 2. The module can be described as Eq. 1:

$$
\begin{aligned}
F &= \left\{ f_{occasion}, f_{season}, f_{age}, f_{ethnicity}, f_{body} \right\}, \\
W_f &= Sigmoid\left(\frac{1}{|F|-1} \sum_{x \in F-f} f \otimes x^T \right), \\
F_{fusion} &= \frac{1}{|F|} \sum_{f \in F} W_f \otimes f,
\end{aligned}
\tag{1}
$$

where f_x refers to the features of data x, $Sigmoid$ refers to the Sigmoid activation function, $\otimes$ is a multiplication operator, and F_{fusion} is an output from the pairwise attention module.

Finally, the F_{fusion} is passed to the fully connected layer and Softmax Layer to predict the probability of each class (i.e., 24 categories for clothing items and 65 types for clothing attributes).

3.3 Query stage

For this stage, the output from the recommendation stage is used to query the best fitting outfits from an outfit dataset. For this, a GloVe embedding [15] is used to embed the output of the recommendation stage into a textual feature

vector. After that, we use the cosine similarity to measure the similarity between recommended output and all available outfits in the query dataset. Finally, the chosen outfit will be the result of the query stage, a set of clothing items, that best matches the output from the recommendation stage. We employ:

$$Similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{N} A_i \times B_i}{\sqrt{\sum_{i=1}^{N} A_i^2} \times \sqrt{\sum_{i=1}^{N} B_i^2}}, \tag{2}$$

where A and B are different feature vectors with the same size and dimension, and N is the dimensionality of features. Moreover, a two-step filtering step is used for the refinement of the results. For this, we choose to filter for attribute information first, then for clothing items second, to query the final outfit.

4 Evaluations

In this section, we evaluate the performance of the proposed method in comparison to existing methods.

4.1 Environment

Datasets. For the recommendation stage, we employ the Street-Fashion-Style [5] (SFS) dataset, which is a collection of street photos. It provides difficult to collect annotations such as suitable outfits for specific events. In each data sample, a user image is connected with an outfit, including relevant information such as appropriate occasion, current season, and details of each clothing item (e.g. category, color, or material). As discussed in Sec. 3.1, we perform a pre-processing step to augment the data for more detailed user information. For this, we use the LightFace Library [19] which is built upon VGG-Face [13] to extract age and ethnicity information. Next, the user's body shape is predicted by FrankMocap [18] which includes SMPL-X [14]. Finally, after removing incomplete or missing data, our pre-processed dataset results in 85,353 data samples. The dataset is split into 70% for training, 15% for testing, and 15% for validation. Examples of the pre-processed data are shown in Fig. 3.

For the query stage, we employ the Polyvore [6] dataset. It contains 164,000 clothing items which group into 21,889 outfits. Each clothing item is further annotated with category, style, and details (e.g. brand, color, or material).

Ground-truth. After the dataset is processed, we define a *ground-truth* used for evaluation. For each combination of inputs, we collect all data samples fitting this scenario. Next, we determine the likelihood distribution of which clothing types and styles are the most fitting, essentially summarizing the outfit choices of exiting users. With this, we gain 2,545 scenarios across the 85,353 data samples with a likelihood distribution across 24 types and 65 styles of clothing.

USER_NAME	PIC_NAME	IMAGE	OCCASION	SEASON	AGE	ETHNICITY	BODY SHAPE IMAGE
29Skirts	91006.jpg		Vacation	Spring	18-27	White	
fashionophile	123139.jpg		Everyday	Fall	28-37	Asian	

Fig. 3. Example of preprocessed dataset. It is based on the Street-Fashion-Style [5] dataset, but includes extra data extracted through our preprocessing step.

Comparison methods. To evaluate our framework we use two comparison methods. First, we implemented a naïve baseline model. It is just convolution layers with an Attentional Feature Fusion (AFF) [4] to fuse all input features at the same time. Second, we use ViBE [10] as an existing comparison method.

Proposed method. We implemented the proposed method as introduced in this paper. For the recommendation stage, we embed each input into a 64-dim feature vector. Next, the feature extractor will transform it into a 1024-dim feature vector. The pairwise attention module combines the 5×1024-dim features into a 1024-dim feature vector. Lastly, it is passed into the fully connected layer and softmax layer to map into a 24-dim output for clothing items and 65-dim output for clothing attributes. It is trained using Adam optimizer [11] with an initial learning rate of 0.005, decay learning rate with gamma(γ) = 0.1 every 7 steps, and trained for 20 epoches. For the query stage, we create a 300-dim GloVe [15] vector to embed the clothing items and attribute from the recommendation model as well as the information of outfits in the Polyvore dataset. For top-k query, k is set to 5 and similarity threshold = 0.5.

4.2 Experiments

Quantitative evaluation. For this experiment, we analyze the quantitative metrics MSE, mAP@k, and mAR@k and compare our proposed method to the two comparison methods. We predict the recommended output of 24 clothing types and 65 clothing styles using each method. Then, we compare the output from each method with our previously defined ground-truth likelihood. To better understand the performance for different choices of user inputs, we also evaluate sub-models using only a subset of personal attributes for the recommendation.

Subjective evaluation. To evaluate the human perception of recommended outfits, we do a subjective evaluation by a user study. In a questionnaire, we asked participants to decide between the two outfits for a given query. Each question

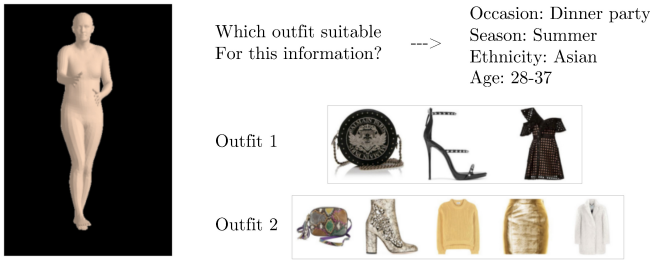

Fig. 4. Example of question in user study. Each participant is asked to decide the better outfit for a certain input query, as shown in the top right. The selectable outfits 1 and 2 are recommendations generated by the comparison and proposed method, respectively.

shows attributes such as body shape, occasion, season, age, and ethnicity, and then two outfit choices. The two outfits are generated by ViBE [10] and the proposed method, making it possible to evaluate which method's outputs is preferred by each participant. An example of the survey is shown in Fig. 4. We gathered 43 outfit pairs made from a random season, occasion, and age. The survey had 34 participants of Asian ethnicity, which can be divided into two genders, 24 female and 10 male. As all participants were of Asian ethnicity, all queries were done with Asian ethnicity.

4.3 Results

Quantitative evaluation. First, we evaluate the performance of the proposed method, as shown in Table 1 (item recommendation) and Table 2 (attribute recommendation). For item recommendation, the proposed method with all inputs achieves the highest performance in MSE metric, mAP, and mAR at k=5, 20. The proposed method improves 53.29% over the ViBE comparison method, and there is a significant improvement for mAP and mAR at k=5. Note, that the approach proposed by ViBE is mostly targeting body-type recommendations, unlike ours which covers a full range of user- and environmental attributes. For attribute recommendation, the ViBE model does not recommend clothing attributes. Because of this, we cannot compare to ViBE, but only to the naïve baseline as our proposed method with several different input settings. The proposed method with all input information has an average performance better than every other model.

Ablation study on fusion method. When preparing the dataset, we noticed an imbalance of annotations as well as a long-tailed distribution which would yield issues with a high-dimensional input recommendation. To test this, we also evaluated the naïve baseline method, as it uses no comprehensive pairwise attention to solve this dataset issue. As expected, the results shown in Table 3 prove this intuition, by showing a decreasing performance in mAP@5 and mAP@10 when

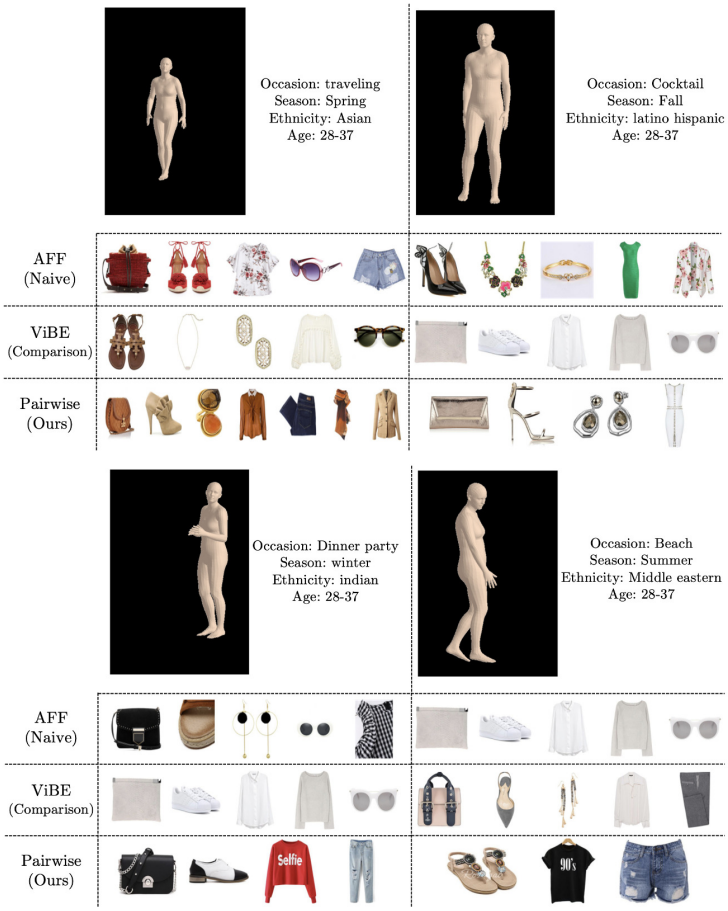

Fig. 5. Examples of outfit recommendation, comparing the proposed method to the comparison methods on four different queries.

Table 1. Quantitative results on clothing item recommendation comparisons of the proposed method with comparison method. The input for our method is abbreviated as **(O)**ccasion, **(S)**eason, **(A)**ge, **(E)**thnicity, and **(B)**ody shape.

Model input	MSE	mAP@5	mAR@5	mAP@20	mAR@20
Comp. method (ViBE [10])	0.00676	0.4859	0.4865	0.7103	0.6108
Naïve (AFF, All)	0.30336	0.5708	0.5714	0.8165	0.8676
Ours (O+S)	0.00032	0.8023	0.8029	0.8781	0.8814
Ours (O+S+A)	0.00032	0.8039	0.8045	0.8773	0.8822
Ours (O+S+A+E)	**0.00030**	0.8279	0.8286	0.8893	0.8900
Ours-Proposed (All)	**0.00030**	**0.8311**	**0.8316**	**0.8907**	**0.8905**

Table 2. Quantitative results on clothing attribute recommendation comparisons of the proposed method with comparison method. The input for our method is abbreviated as **(O)**ccasion, **(S)**eason, **(A)**ge, **(E)**thnicity, and **(B)**ody shape.

Model input	MSE	mAP@5	mAR@5	mAP@20	mAR@20
Naïve (AFF, All)	0.00513	0.7427	0.5701	0.7810	0.4356
Ours (O+S)	0.00016	0.8234	0.8240	0.8057	0.8113
Ours (O+S+A)	0.00027	0.7849	0.7854	0.7991	0.7789
Ours (O+S+A+E)	0.00592	**0.8842**	0.6263	**0.9459**	0.2871
Ours-Proposed (All)	**0.00015**	0.8377	**0.8382**	0.8188	**0.8203**

Table 3. Quantitative results of naïve method. The input dimension is abbreviated as **(O)**ccasion, **(S)**eason, **(A)**ge, **(E)**thnicity, and **(B)**ody shape.

Model input	mAP@5	mAP@10	mAP@15	mAP@20
Naïve (O+S)	**0.7137**	**0.7850**	0.8052	0.8136
Naïve (O+S+A)	0.6490	0.7568	0.7939	0.8113
Naïve (O+S+A+E)	0.5597	0.7235	0.7961	0.8311
Naïve (All)	0.5706	0.7331	**0.8063**	**0.8419**

Table 4. Ablation study on the attention fusion model, comparing the naïve method and the proposed method.

Fusion method	mAP@5	mAP@10	mAP@15	mAP@20
Naïve (AFF)	0.5708	0.5714	0.8165	0.8676
Ours (Pairwise)	**0.8311**	**0.8316**	**0.8907**	**0.8905**

Table 5. Subjective evaluation. Questionnaire result comparisons of the proposed method with comparison method.

Preferred recommendation	Female	Male	All
From ViBE [10]	**19** (44.19%)	15 (34.88%)	20 (46.51%)
From Proposed model	**19** (44.19%)	**21** (48.84%)	**22** (51.16%)
Tied between both	5 (11.63%)	7 (16.28%)	1 (2.33%)

adding extra input features. To further ablate this, we compared the performance of attention fusion, shown in Table. 4. It can be seen that by adding the pairwise attention, the performance increases around 45.6% over the naïve baseline method at mAP@5 and mAP@10.

Subjective evaluation The results of the subjective evaluation method are shown in Table 5. Out of 43 outfits, the proposed model gave the better recommendation for 22 outfits, giving slightly better recommendations than the comparison method. While these results are close, a tendency towards the proposed model can be seen. It is larger for male participants, where a majority of users preferred our recommendation. For female participants, the results are tied.

Qualitative evaluation Fig. 5 shows examples of outfit recommendations for each tested model with all input information. Each example shows the subject's estimated body shape and personal attributes as an input for each model, and each row shows the output for the corresponding model. The proposed method can generally recommend outfits based on some specific information such as season and occasion. For example, in the first column, the proposed method suggests a comfortable outfit with a colorful long sleeve t-shirt and jeans for dinner parties in the winter season. In the second column, the proposed method recommends a t-shirt, short jeans, and sandals for going to the beach in summer which might be more suitable than the recommended outfit from the comparison method that suggests a blouse, shirt, and heels.

5 Conclusion

In this work, we proposed a novel method for recommending clothing items and styles with high-dimensional personalization, including personal attributes (age, ethnicity, body shape) and environment information (occasion and season). To solve data imbalance issues with existing datasets, we introduce a pairwise attention module to improve the performance of the recommendation. This module can weigh the importance between each input data type, better understanding the data in case of lesser-known input combinations. We evaluate our proposed method and compare it to an existing method. We can show an average improvement in 53.29% and 38.24% on recommending clothing items and styles respectively over the comparison method. The ablation study on attention methods confirms dataset imbalance issues. Moreover, in a subjective evaluation with human participants, we can show a tendency towards preferring our recommendations over the comparison method.

References

1. Amed, I., Balchandani, A., Berg, A., Hedrich, S., Jensen, J.E., Rölkens, F.: The State of Fashion 2021. McKinsey & Company (2021)
2. Amed, I., Berg, A., Balchandani, A., Hedrich, S., Rolkens, F., Young, R., Ekelof, J.: The state of fashion 2020. Business of Fashion and McKinsey & Company (2020)
3. Amed, I., Bergsara, A., Kappelmark, S., Hedrich, S., Andersson, J., Young, R., Drageset, M.: The state of fashion 2019; business of fashion, mckinsey & company. 2019 (2018)
4. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: IEEE Winter Conf. Appl. Computer Vision, WACV. pp. 3559–3568 (2021). https://doi.org/10.1109/WACV48630.2021.00360
5. Gu, X., Wong, Y., Peng, P., Shou, L., Chen, G., Kankanhalli, M.S.: Understanding fashion trends from street photos via neighbor-constrained embedding learning. In: Proc. 2017 ACM Multimedia Conf. pp. 190–198 (2017). https://doi.org/10.1145/3123266.3123441
6. Han, X., Wu, Z., Jiang, Y., Davis, L.S.: Learning fashion compatibility with bidirectional lstms. In: Proc. 2017 ACM Multimedia Conf. pp. 1078–1086 (2017). https://doi.org/10.1145/3123266.3123394

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conf. Computer Vision and Pattern Recognition, CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
8. He, X., Du, X., Wang, X., Tian, F., Tang, J., Chua, T.: Outer product-based neural collaborative filtering. In: Lang, J. (ed.) Proc. Twenty-Seventh Int. Joint Conference on Artificial Intelligence, IJCAI. pp. 2227–2233. ijcai.org (2018)
9. Hidayati, S.C., Hsu, C., Chang, Y., Hua, K., Fu, J., Cheng, W.: What dress fits me best?: Fashion recommendation on the clothing style for personal body shape. In: 2018 ACM Multimedia Conf. pp. 438–446 (2018). https://doi.org/10.1145/3240508.3240546
10. Hsiao, W., Grauman, K.: Vibe: Dressing for diverse body shapes. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 11056–11066 (2020). https://doi.org/10.1109/CVPR42600.2020.01107
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd Int. Conf. Learning Representations, ICLR (2015)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: 26th Ann. Conf. Neural Information Processing Systems (NIPS). pp. 1106–1114 (2012)
13. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proc. British Machine Vision Conference BMVC. pp. 41.1–41.12 (2015). https://doi.org/10.5244/C.29.41
14. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE Conf. Computer Vision and Pattern Recognition, CVPR. pp. 10975–10985. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01123
15. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proc. 2014 Conf. Empirical Methods in Natural Language Processing, EMNLP. pp. 1532–1543 (2014). https://doi.org/10.3115/v1/d14-1162
16. Rawat, Y.S., Kankanhalli, M.S.: Contagnet: Exploiting user context for image tag recommendation. In: Proc. 2016 ACM Conf. Multimedia. pp. 1102–1106. ACM (2016)
17. Rendle, S.: Factorization machines with libfm. ACM Trans. Intell. Syst. Technol. **3**(3), 57:1–57:22 (2012)
18. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. CoRR arXiv:2008.08324 (2020)
19. Serengil, S.I., Ozpinar, A.: Lightface: A hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). pp. 1–5 (2020)
20. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
21. Verma, D., Gulati, K., Goel, V., Shah, R.R.: Fashionist: Personalising outfit recommendation for cold-start scenarios. In: 28th ACM Int. Conf. Multimedia. pp. 4527–4529 (2020). https://doi.org/10.1145/3394171.3414446
22. Zheng, L., Noroozi, V., Yu, P.S.: Joint deep modeling of users and items using reviews for recommendation. In: Proc. Tenth ACM Int. Conf. Web Search and Data Mining, WSDM. pp. 425–434 (2017)