Modeling Fashion Compatibility with Explanation by using Bidirectional LSTM

Pang Kaicheng[1], Zou Xingxing[2], Wai Keung Wong[1, 2, *]
[1]Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong
[2]Laboratory for Artificial Intelligence in Design, Hong Kong
{kaicpang.pang, aemika.zou}@connect.polyu.hk, calvin.wong@polyu.edu.hk

Abstract

The goal of this paper is to model the fashion compatibility of an outfit and provide the explanations. We first extract features of all attributes of all items via convolutional neural networks, and then train the bidirectional Long Short-term Memory (Bi-LSTM) model to learn the compatibility of an outfit by treating these attribute features as a sequence. Gradient penalty regularization is exploited for training inter-factor compatibility net which is used to compute the loss for judgment and provide its explanation which is generated from the recognized reasons related to the judgment. To train and evaluate the proposed approach, we expanded the EVALUATION3 dataset in terms of the number of items and attributes. Experiment results show that our approach can successfully evaluate compatibility with reason.

1. Introduction

Outfit recommendation is a good way for online shopping platform to do cross-selling. Thus, fashion compatibility learning attracts many attention for its huge potential economic value [3].

Mainstream method is adopting metric learning [8, 6, 15, 5, 10, 16, 11, 12, 17]. Items of an outfit are transformed into embeddings and the metric distance will be calculated. [5] learned the compatibility among fashion items based on the Bi-LSTM network which assumes the outfit as a sequence input. Other studies use the Conditional Random Field [11], and clothing style modeling [16, 1] to estimate the fashion compatibility.

Many studies have been devoted to provide explanations [4, 14, 13]. In particular, [7] provided fashion suggestions and generated abstract comments as explanations at the same time. [18] introduced Visual and Textual Jointly Enhanced Interpretable model to generate the interpretable fashion recommendations. [2] proposed a co-attentive multi-task learning model to generate an explainable recommendation. However, these approaches use too many comments or reviews from the social network users that make the training data lacking evaluation from fashion experts. This training manner leads to the inconclusive explanation results, and are weak on giving an explanation which closely related to the judgment predicted by the model.

In this work, we propose an outfit compatibility evaluation framework which provides a closely explanation with the predicted judgment. An outfit comprises of multiple different items is considered as a sequence. The judgment and its reason are trained jointly using the Bi-LSTM network. The overview of the system is illustrated in Figure 1. We inherited the classification pattern of judgment from this work [19]. The compatibility is divided into three levels, namely *good*, *normal*, and *bad*. Our main contributions can be summarized as follows: 1. We employ the bidirectional LSTM to deal with the problem of fashion compatibility learning with *variable-length items* while *providing a convincing reason* for the judgment via the gradient penalty. 2. We expand the EVALUATION3 dataset with more comprehensive outfit (*i.e.* each outfit comprises multiple items), attributes, judgments, and reasons. 3. We demonstrate the practical value of our work through the experiments and a demo website based on our proposed approach. The demo website can evaluate the compatibility of an outfit and also provide a comprehensible explanation sentence in seconds.

2. Approach

The pipeline of the compatibility network is shown in Figure 1. We denote the judgment set by $\mathcal{J} = \{good, normal, bad\}$, and the reason set by $\mathcal{R} = \{color, print, material, \cdots, shape\}$. Note that the element of reason set is an aggregate (e.g. *color* represent color of top, color of bottom, *etc*.).

2.1. Bidirectional LSTM Architecture

For the item features extracted stage, we use the bidirectional LSTM network to learn the compatibility of an outfit,

[*] Corresponding author

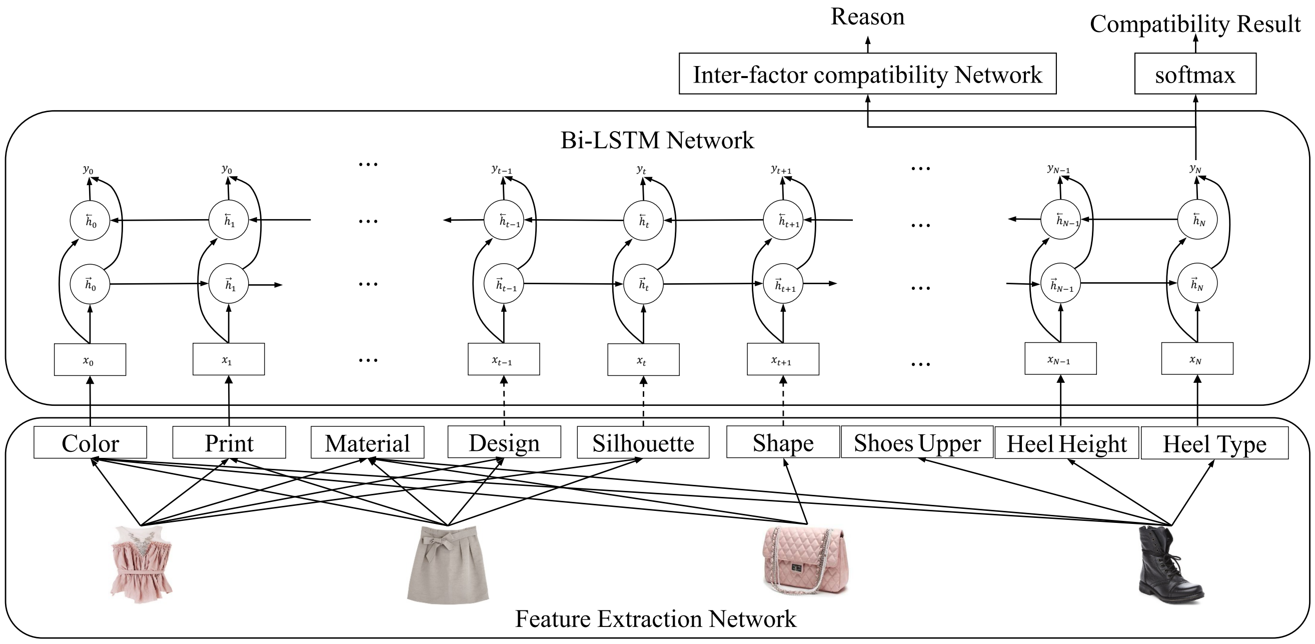


Figure 1. The pipeline of fashion compatibility network. The bidirectional LSTM takes the last 512-dimensional feature maps from the CNNs as input and each feature is considered as a contributing factor. A softmax layer is used to map the output of Bi-LSTM's last time step to the compatibility judgment space. The inter-factor compatibility network will evaluate the corresponding reason for judgment. It takes the output of Bi-LSTM and features as inputs and uses gradient penalty to learn the reason.

because the standard RNN network has two main shortcomings that make it unsuitable for outfit compatibility learning. 1. RNN network cannot process very long sequences if we uses tanh activation as its activation function and thus is not able to keep track of long-term dependencies neither. 2. Our input data is the embedding features of an outfit, and thus there is no reason not to exploit future item features as well.

As illustrated in Bi-LSTM stage of Figure 1, the Bi-LSTM Network computes the *forward* hidden sequence $\overrightarrow{h}$, the *backward* hidden sequence $\overleftarrow{h}$ and the output sequence y by iterating the backward layer from $t = T$ to 1, the forward layer from $t = 1$ to T and then updating the output layer:

$$\overrightarrow{h}_t = \mathcal{H}(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \tag{1}$$

$$\overleftarrow{h}_t = \mathcal{H}(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{3}$$

where $W_{\alpha\beta}$ is the weight matrix between vector α and β, b represent the bias term. $\mathcal{H}$ is the Long Short-Term Memory (LSTM) cell.

In this work, we uses a softmax layer to define a separate output distribution $\Pr(k|t)$ at each step t along the input sequence as follows:

$$\begin{aligned} y_t &= W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \\ \Pr(k|t) &= \frac{\exp(y_t[k])}{\sum_{k'=1}^{K} \exp(y_t[k'])} \end{aligned} \tag{4}$$

where k is the number of the judgments, and $y_t[k]$ is the k-th element of the output vector y_t. The loss for a given outfit F can be calculated as:

$$L_{\text{judgment}} = -\frac{1}{K} \sum_{k=1}^{K} \log\Pr(k|t) \tag{5}$$

2.2. Gradient Penalty Architecture

The neuron importance weight α_k^c defined in [9] is:

$$\alpha_k^c := \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{6}$$

where i, j iterates over the spatial dimensions and Z is the number of pixels in the feature map. We perform a weighted product of forward activation maps, and follow it by a ReLU to obtain the heatmap $H_c^{\text{Grad-CAM}}$:

$$H_c^{\text{Grad-CAM}} := \text{ReLU}(\sum_k \alpha_k^c A^k) \tag{7}$$

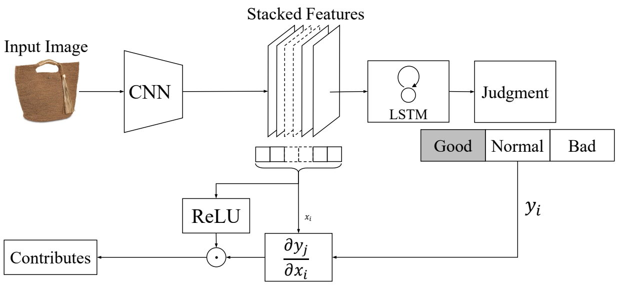


Figure 2. Inter-factor compatibility network overview: given the compatibility output computed by Bi-LSTM and the stacked features as input. We pointwise multiply the gradients via backprop with the stacked feature to obtain the contribution of each element for the decision of judgment.

Following this work [19], we exploit gradient penalty to predict the reason for judgment. As shown in Figure 2, we define the *contribution* of each element for the decision of judgment as contrib_j:

$$\text{contrib}_j(x_i) := \frac{\partial y_i}{\partial x_i} \odot \text{ReLU}(x_i) \tag{8}$$

where y_i is the logit for the judgment $j \in \mathcal{J}$, and x_i is one element of compatibility feature $x_i \in \mathbf{x}$.

The *positive contribution* of x_i for the judgment is:

$$\text{contrib}_j^+(r) := \frac{1}{|I_r|} \sum_{i \in I_r} \text{ReLU}(\frac{\partial y_i}{\partial x_i}) \odot \text{ReLU}(x_i) \tag{9}$$

where I_r is the index set of neurons for factor $r \in \mathcal{R}$.

We train the network with specially designed regularizations so that the main reason predicted by the network is aligned with pre-labeled data. Cross-entropy regularizer is used to compute the reason loss for training. The mathematical form is:

$$F_r := \sum_{j \in \mathcal{J}} \mathbb{1}_{j^{gt}}(j) \cdot \text{contrib}_j^+(r) - \text{contrib}_{\text{normal}}^+(r) \tag{10}$$

$$L_{\text{reason}} = -\log(\frac{\exp(F_{r^{gt}})}{\sum_{r \in R} \exp(F_r)}) \tag{11}$$

where $\mathbb{1}_{j^{gt}}$ is an indicator function for ground-truth judgment. If judgment j is the same as ground-truth label, $\mathbb{1}_{j^{gt}} = 1$; else, $\mathbb{1}_{j^{gt}} = 0$. The total loss L is described as follows:

$$L = L_{\text{judgment}} + \alpha L_{\text{reason}} \tag{12}$$

where α is a hyper-parameter which is used to control the effect of reason regularization. We jointly train the Bi-LSTM network and inter-factor network, as shown in the definition of contribution (Equation 9) and reason (Equation 10), because the loss term L_{reason} penalizes the gradient which directly affects network parameters.

Methods	judgment accuracy	reason accuracy
Multi-CLS-Part	74.4 ± 0.5	74.8 ± 3.1
IFIV[14]	73.3 ± 0.7	35.9 ± 4.5
Reason linear	72.8 ± 1.3	68.3 ± 2.4
Reason square	72.1 ± 1.6	73.8 ± 1.6
Reason cross-entropy	74.8 ± 1.7	76.7 ± 3.6
Bi-LSTM (Ours)	72.0 ± 0.01	**77.6 ± 3.9**

Table 1. Comparison of different methods on the updated EVALUATION3 test set. All the evaluating experiments are repeated 6 times, and the values after ± are the mean square error.

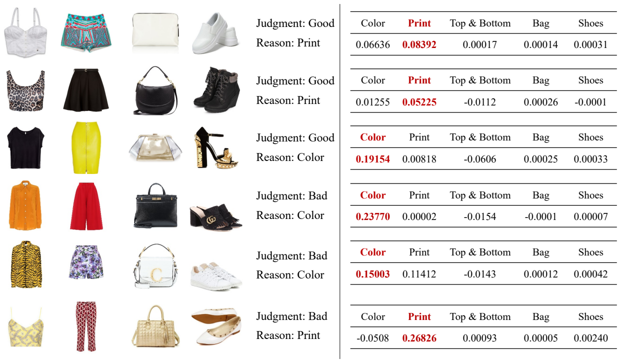
3. Experiment

Data Construction. The Polyvore fashion website is a popular fashion website. Zou *et al.* [19] presented a dataset named EVALUATION3 whose image source is a subset of the *Polyvore dataset* [5]. However there are two problems with the EVALUATION3 dataset. Firstly, it lacks some attributes of bags and shoes. Secondly, due to the number of items included in an outfit is increased, from two items to multiple items, the entire original evaluation result is not applicable to the current outfit structure. To this end, we labeled the corresponding attributes of bags and shoes. To address the second problem, all labels in the EVALUATION3 dataset have been manually annotated from scratch.

To summarize our dataset, there are 34,479 outfits which are split into 29,479 for training, 3,000 for validation, and 2,000 for testing. Each outfit comprises a top, a bottom, a bag, and a pair of shoes. Each outfit has a corresponding judgment label with a reason label. If an outfit belongs to the normal level, the reason label will leave blank.

Training details. We use Resnet18 to learn the embedding features of each attribute. Each of them is extracted from the Resnet18 model with different parameters. These networks are optimized by using Adam method with an initial learning rate of 0.001 in conjunction with the weight decay of 5×10^{-5}. The learning rate will be divided by 10 every 10 epochs after 30 epochs. For Bi-LSTM, we use the SGD optimization method with an initial learning rate of 0.001, and weight decay of 5×10^{-4} for 140 epochs. We use one layer Bi-LSTM with the hidden layer dimension of 500 as our compatibility evaluation model architecture. The learning rate will be divided by 10 after 84 epochs. We use cross-entropy as the form of regularization in terms of learning the corresponding reason.

Quantitative analysis. The goal of our proposed method is to evaluate the compatibility of an outfit that contains multiple items and provide its reason. To demonstrate the effec-



Judgment: Good
Reason: Print

Color	**Print**	Top & Bottom	Bag	Shoes
0.06636	**0.08392**	0.00017	0.00014	0.00031

Judgment: Good
Reason: Print

Color	**Print**	Top & Bottom	Bag	Shoes
0.01255	**0.05225**	-0.0112	0.00026	-0.0001

Judgment: Good
Reason: Color

Color	Print	Top & Bottom	Bag	Shoes
0.19154	0.00818	-0.0606	0.00025	0.00033

Judgment: Bad
Reason: Color

Color	Print	Top & Bottom	Bag	Shoes
0.23770	0.00002	-0.0154	-0.0001	0.00007

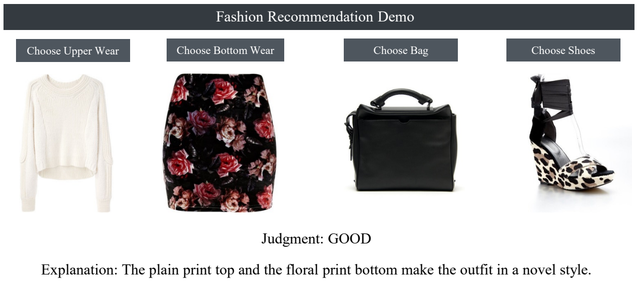
Judgment: Bad
Reason: Color

Color	Print	Top & Bottom	Bag	Shoes
0.15003	0.11412	-0.0143	0.00012	0.00042

Judgment: Bad
Reason: Print

Color	**Print**	Top & Bottom	Bag	Shoes
-0.0508	**0.26826**	0.00093	0.00005	0.00240

Figure 3. Qualitative analysis of the proposed approach. There are six outfits in this figure and each outfit contains four different items. On the left side of the vertical line shows the judgment and its reason predicted via the model. The contribution values of different candidate reasons are given on the right side of the vertical line and the maximum values are marked in red.



Judgment: GOOD

Explanation: The plain print top and the floral print bottom make the outfit in a novel style.

Figure 4. We show a website demo application that can predict the compatibility of an outfit and give the corresponding explanation.

tiveness of our approach, we compare with the regularizing reason method [19] in terms of the *judgment accuracy* and *reason accuracy*. *Judgment accuracy* is calculated by dividing the number of correct predicted judgments by the total size of the test dataset. We calculate the ratio of the number of correct predicted reasons to the number of correct predicted judgments as the *reason accuracy*. It should be noted that when calculating the *reason accuracy*, the number of correct predicted judgment only includes the predicted values of *good* and *bad*, and the predicted values of *normal* is excluded. This is because our model will not give a reason if the predicted judgment value is *normal*.

The evaluation result is shown in Table 3. The method in the first row of Table 3 is a multi-task classification model, which is to classify judgments and reasons separately. The Item Feature Influence Value (IFIV), the method in the second row, did not learn the reason with supervision. We can see from the table that the *judgment accuracy* and *reason accuracy* of our approach are 72.0 ± 0.01 and 77.6 ± 3.9 respectively. In terms of the *reason accuracy*, our method outperforms other methods. In regards to for *judgment accuracy*, the regularizing reason method achieves 2.8 percentage higher than our method.

Qualitative analysis. We also show some evaluation results of outfits with numerical details in Figure 3. In order to facilitate the demonstration of qualitative analysis, we predict the compatibility of an outfit that contains four items. In fact, our model can evaluate the compatibility of an outfit that contains multiple different items. Take the first line of Figure 3 as an example. The text on the left side of the vertical line indicates that the current outfit compatibility is good, and its reason is *print*. The table on the right side of the vertical line indicates the weight values of each reason contribution calculated by our model. The weight score of *print* is the largest (red mark) among these five candidate reasons. This numerical result proves that our model is consistent with the ground truth.

The ground truth of compatibility in the fourth line is *bad* which is consistent with our visual perception. This feeling is mainly caused by the coat in orange and the pants in red. Our data can also reflect this situation. You can see that there is a huge difference between **0.23770** in color and **0.00002** in print. We can also find this situation in the sixth line, but this time the weight score of the print (**0.26826**) is larger than color (**-0.0508**). This huge numerical difference shows that our model indeed learned how to predict the corresponding reason for its corresponding judgment.

In addition, we developed a website demo application based on this compatibility evaluation system as shown in Figure 4. The users need to upload four different pictures which are the top, bottom, bag, and shoes to this website, and the compatibility of this outfit and the explanation generated according to its reason will be provided. The compatibility of the given outfit belongs to *good*, and its corresponding reason is "*The plain print top and the floral print bottom make the outfit in a novel style*". Recall that we first extract all the color, print, and attributes features of items. The explanation sentence is composed of features corresponding to its reason. For this example, the reason why the model classifies the outfit as *good* is print. So we use the print feature extraction network to classify the print type of the top and bottom, which are *plain* print and *floral* print in this case. Finally, we use the pre-designed sentence templates to generate the explanation.

4. Conclusion

In this work, we present a fashion compatibility evaluation system which is achieved by jointly training a Bi-LSTM model and an inter-factor compatibility network. The proposed approach has a great potential to be practically applied in the fashion retail industry.

5. Acknowledgement

This research is funded by the Laboratory for Artificial Intelligence in Design (Project Code: RP3-1), Hong Kong.

References

[1] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[2] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2137–2143. AAAI Press, 2019.

[3] Xiaoling Gu, Fei Gao, Min Tan, and Pai Peng. Fashion analysis and understanding with artificial intelligence. *Information Processing & Management*, 57(5):102276, 2020.

[4] Xianjing Han, Xuemeng Song, Jianhua Yin, Yinglong Wang, and Liqiang Nie. Prototype-guided attribute-wise interpretable scheme for clothing matching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 785–794, 2019.

[5] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017.

[6] Wei-Lin Hsiao and Kristen Grauman. Learning the latent ”look”: Unsupervised discovery of a style-coherent embedding from fashion images. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Explainable fashion recommendation with joint outfit matching and comment generation. *arXiv preprint arXiv:1806.08977*, 2, 2018.

[8] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, 2015.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[10] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. Compatibility family learning for item recommendation and generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] 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.

[12] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 753–761, 2017.

[13] Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In *Proceedings of The Web Conference 2020*, pages 837–847, 2020.

[14] Pongsate Tangseng and Takayuki Okatani. Toward explainable fashion recommendation. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[15] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

[16] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015.

[17] Xin Wang, Bo Wu, and Yueqi Zhong. Outfit compatibility prediction and diagnosis with multi-layered comparison network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 329–337, 2019.

[18] Qianqian Wu, Pengpeng Zhao, and Zhiming Cui. Visual and textual jointly enhanced interpretable fashion recommendation. *IEEE Access*, 8:68736–68746, 2020.

[19] Xingxing Zou, Zhizhong Li, Ke Bai, Dahua Lin, and Waikeung Wong. Regularizing reasons for outfit evaluation with gradient penalty. *arXiv preprint arXiv:2002.00460*, 2020.