Modeling Fashion Compatibility with Explanation by using Bidirectional LSTM

Pang Kaicheng\textsuperscript{1}, Zou Xingxing\textsuperscript{2}, Wai Keung Wong\textsuperscript{1, 2}; *  
\textsuperscript{1}Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong  
\textsuperscript{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 $\mathcal{J} = \{\text{good}, \text{normal}, \text{bad}\}$, and the reason set by $\mathcal{R} = \{\text{color}, \text{print}, \text{material}, \ldots, \text{shape}\}$. 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,
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 use 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:

\[
\overrightarrow{h}_t = \mathcal{H}(W_{xh} \overrightarrow{x}_t + W_{hh} \overrightarrow{h}_{t-1} + b) \\
\overleftarrow{h}_t = \mathcal{H}(W_{xh} \overleftarrow{x}_t + W_{hh} \overleftarrow{h}_{t+1} + b) \\
y_t = W_{hy} \overrightarrow{h}_t + W_{hy} \overleftarrow{h}_t + b
\]

where \( W_{\alpha\beta} \) is the weight matrix between vector \( \alpha \) and \( \beta \), \( b \) 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 \( \text{Pr}(k|t) \) at each step \( t \) along the input sequence as follows:

\[
y_t = W_{hy} \overrightarrow{h}_t + W_{hy} \overleftarrow{h}_t + b \\
\text{Pr}(k|t) = \frac{\exp(y_t[k])}{\sum_{k'=1}^{K} \exp(y_t[k'])}
\]

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:

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

### 2.2. Gradient Penalty Architecture

The neuron importance weight \( a^c_k \) defined in [9] is:

\[
a^c_k := \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c_i}{\partial A^k_{ij}}
\]

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 \( H^{\text{Grad-CAM}}_c \):

\[
H^{\text{Grad-CAM}}_c := \text{ReLU}(\sum_k a^c_k A^k)
\]
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:

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

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:

\[
F_r := \sum_{j \in \mathcal{J}} 1_{j \neq t}(j) \cdot \text{contrib}^+_j(r) - \text{contrib}^+_\text{normal}(r) \tag{10}
\]

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

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

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

where \(\alpha\) 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 table below.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Judgment accuracy</th>
<th>Reason accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-CLS-Part</td>
<td>74.4 ± 0.5</td>
<td>74.8 ± 3.1</td>
</tr>
<tr>
<td>IFIV[14]</td>
<td>73.3 ± 0.7</td>
<td>35.9 ± 4.5</td>
</tr>
<tr>
<td>Reason linear</td>
<td>72.8 ± 1.3</td>
<td>68.3 ± 2.4</td>
</tr>
<tr>
<td>Reason square</td>
<td>72.1 ± 1.6</td>
<td>73.8 ± 1.6</td>
</tr>
<tr>
<td>Reason cross-entropy</td>
<td>74.8 ± 1.7</td>
<td>76.7 ± 3.6</td>
</tr>
<tr>
<td>Bi-LSTM (Ours)</td>
<td>72.0 ± 0.01</td>
<td>77.6 ± 3.9</td>
</tr>
</tbody>
</table>

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.
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.

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


date 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


