Learning the Compositional Visual Coherence for Complementary Recommendations

Zhi Li¹, Bo Wu², Qi Liu¹,³, Likang Wu³, Hongke Zhao⁴ and Tao Mei⁵

¹Anhui Province Key Laboratory of Big Data Analysis and Application, School of Data Science, University of Science and Technology of China, ²Columbia University, ³School of Computer Science and Technology, University of Science and Technology of China, ⁴The College of Management and Economics, Tianjin University, ⁵JD AI Research

{zhili03, wulk}@mail.ustc.edu.cn, bo.wu@columbia.edu, qiliuql@ustc.edu.cn, hongke@tju.edu.cn, tmei@jd.com

Abstract

Complementary recommendations, which aim at providing users product suggestions that are supplementary and compatible with their obtained items, have become a hot topic in both academia and industry in recent years. Existing work mainly focused on modeling the co-purchased relations between two items, but the compositional associations of item collections are largely unexplored. Actually, when a user chooses the complementary items for the purchased products, it is intuitive that she will consider the visual semantic coherence (such as color collocations, texture compatibilities) in addition to global impressions. Towards this end, in this paper, we propose a novel Content Attentive Neural Network (CANN) to model the comprehensive compositional coherence on both global contents and semantic contents. Specifically, we first propose a Global Coherence Learning (GCL) module based on multi-heads attention to model the global compositional coherence. Then, we generate the semantic-focal representations from different semantic regions and design a Focal Coherence Learning (FCL) module to learn the focal compositional coherence from different semantic-focal representations. Finally, we optimize the CANN in a novel compositional optimization strategy. Extensive experiments on the large-scale real-world data clearly demonstrate the effectiveness of CANN compared with several state-of-the-art methods.

1 Introduction

Recommender systems are those techniques that support users in the various decision-making process and catch their interest among the overloaded information. For enhancing user satisfaction and recommendation performances, it is an indispensable part to understand how products relate to each other in recommender systems [McAuley et al., 2015a; Li et al., 2018]. Along this line, complementary recommendations [Yu et al., 2019], which aim at exploring item compatible associations to enhance the qualities of each item or another, have become a hot topic in both academia and industry in recent years.

In the literature, previous researches can be clustered into two groups, i.e., unsupervised methods [Tan et al., 2004; Zheng et al., 2009] and supervised models [Zhao et al., 2017; He et al., 2016]. For a long time, researchers mainly used unsupervised methods to model the association rules in the recommendation process [Tan et al., 2004]. Meanwhile, some studies proposed supervised approaches to learn complementary relationships, such as co-purchased [Zhao et al., 2017] and content compatibility [He et al., 2016]. Recently, many researchers attempted to mine the compatibility of fashion items to better understand the complementary relationships in the clothing recommendations [Han et al., 2017; Hsiao and Grauman, 2018]. In spite of the importance of existing studies, the exploration of compositional associations in the complementary item recommendations is still limited.

As a matter of fact, when a user chooses the complementary items for the purchased products, it is intuitive that she will consider the visual coherence (such as color collocations, texture compatibilities) in addition to global impressions. For example, Figure 1 shows the case of a complementary recommendation process. If we only consider co-purchased relations, we may recommend a creamy-white shirt to the user as shown in Figure 1(a). That is because the shirt was co-purchased respectively with three query items by other users. When we take consideration of the compatible relations as shown in Figure 1(b), we can find the missing component of the user’s purchased collection is a pair of shoes. Then, a pair of black booties may be a suitable recommendation. However, for considering compositional relationships and vi-
sual semantic, we can find the mini skirt and shoulder bag are in blue. Therefore, as Figure 1(c) shows, a pair of blue leather pumps is the best-matching. From this example, we can conclude that a good complementary recommender system should model the comprehensive compositional relationships on both global contents and semantic contents. Unfortunately, the exploration of this compositional coherence in the complementary recommendations is still limited.

In this paper, we provide a focused study of the compositional coherence in item visual contents from two perspectives, i.e., the global visual coherence and the semantic-focal coherence. Along this line, we propose a novel Content Attentive Neural network (CANN) to address the complementary recommendations. More specifically, we first propose a Global Coherence Learning (GCL) module based on multi-heads attention to model the global compositional coherence. Then, considering the importance of visual semantics (such as color, texture), we generate the content semantic-focal representations of color-focal, texture-focal and hybrid-focal contents, respectively. Next, we design a Focal Coherence Learning (FCL) module based on a hierarchical attention model to learn the focal coherence from different semantic-focal representations. Finally, we optimize the CANN in a novel compositional optimization strategy. We conduct extensive experiments on a real-world dataset and the experimental results clearly demonstrate the effectiveness of CANN compared with several state-of-the-art methods.

2 Related Works

2.1 Visual Recommendations

With the rapid development of deep learning [Zhao et al., 2019; Wu et al., 2019] and computer vision, visual recommendations have raised lots of interests in both academia and industry and benefited lots of applications, such as image recommendations [McAuley et al., 2015b], movie recommendations [Zhao et al., 2016] and fashion recommendations [Han et al., 2017; Hou et al., 2019]. Some previous works treated visual recommendations as special content-aware recommendations incorporating the visual appearance of the items [He and McAuley, 2016]. Along this line, many researchers directly extracted the visual feature by a pre-trained CNN model and enhanced traditional recommender systems, such as Matrix Factorization [He and McAuley, 2016; Hou et al., 2019]. For example, [He and McAuley, 2016] proposed a recommendation framework to incorporate the visual signal of the items. Recently, many researchers utilized deep learning methods to generate item recommendations [Kang et al., 2017]. To further explore item visual contents, some studies developed deep neural networks to extract the aesthetic information of images [Liu et al., 2017; Yu et al., 2018]. Although various researches have leveraged visual features in recommendation tasks, they usually treated item visual features as side information and the item visual relations have been largely unexploited. In this paper, we propose the compositional visual coherence, which is learning by the attention-based deep neural networks, to deeply model the item complementary relation for recommender systems.

2.2 Complementary Recommendations

For enhancing the recommendations, explicitly modeling the complex relations among items under domain-specific applications is an indispensable part [Liu et al., 2018]. Along this line, many researchers focused on exploring the item combination-effect relations, such as substitutable relations [McAuley et al., 2015a; McAuley et al., 2015b] and complementary relations [Rudolph et al., 2016; Yu et al., 2019]. For a long time, many researchers mainly utilized unsupervised learning methods to explore co-occurrence relations for complementary recommendations [Tan et al., 2004; Zheng et al., 2009]. Recently, more and more studies based on supervised approaches were proposed to model complementary relationships, which were mainly reflected by users’ purchases of the complements [Zhao et al., 2017; He et al., 2016] or item content similarities [McAuley et al., 2015a; Zhang et al., 2018]. Along this line, since natural complementary relationships in the fashion items [Chang et al., 2017; Lo et al., 2019], there was a particular interest in understanding the compatibility [Song et al., 2017; Wang et al., 2019] of fashion items to generate complementary recommendations [Han et al., 2017; Hsiao and Grauman, 2018; Vasileva et al., 2018]. For example, [He et al., 2016] learned the complicated and heterogeneous relationships between items and enhance fashion recommendations. [Han et al., 2017] developed Bi-LSTMs to model the outfit completion process and generate the complementary clothing recommendations. Although these works have considered co-reactions between items, the item compositional coherence in the global and semantic contents cannot be depicted and captured well. In this paper, we propose a focal study on exploring the item compositional relationship of visual content to enhance the complementary item recommendations.

3 CANN: Content Attentive Neural Network

In this section, we introduce our proposed framework for addressing the complementary recommendations. In the real complementary item choosing process, people always want to buy some items that can be compatible with their purchased products. Suppose we have a set of compatible item collections $S = \{O_1, O_2, ..., O_N\}$, and for each collection $O_i = \{p_1, p_2, ..., p_k\}$ contains $k$ compatible items $p$. Meanwhile, we have a scenario that a user has purchased a seed collection of items $P = \{p_1, p_2, ..., p_l\}$, but she is confused to choose complementary items $P^*$ which can make the item collection compatible. To that end, in this paper, we aim to give a complementary suggestion for the target user to help her make the best-matched choice.

Along this line, we propose a content-based attention model, i.e., Content Attentive Neural network (CANN), to address the complementation recommendations. As shown in Figure 2, CANN consists of a Global Coherence Learning (GCL) component and a Focal Coherence Learning (FCL) component to jointly learn the compositional relationships from global and semantic-focal contents. In order to simulate users’ decision-making process on complementary items, we propose a novel compositional optimization strategy to train our proposed CANN.
3.1 Global Coherence Learning

Considering the differences from traditional recommendations, the visual contents of items are an important part of the complementary recommendations e.g., the colors should be compatible and styles should be matching. Along this line, we propose an attention-based module, i.e., Global Coherence Learning (GCL) to learn compositional coherence of item global visual contents.

We develop the Inception-V3 CNN model [Szegedy et al., 2016] as the feature extractor to transform item global visual contents. For each item collection \( P = \{p_1, p_2, ..., p_k\} \), we can generate the items visual representations \( P = \{x_1, x_2, ..., x_k\} \) from the item images, where \( x_i \) is the feature representation derived from the CNN model for the \( i \)-th item. We adopt a fully connected layer for all \( x_i \) to reduce the feature dimension and make the visual content representation fine-tuned in the model training stage. More formally, we generate the final visual representation as following:

\[
x_f^i = \sigma(x_iW^{(1)}_f + b^{(1)}_f)W^{(2)}_f + b^{(2)}_f,
\]

where the weight matrices \( W^{(1)}_f, W^{(2)}_f \) are of shape \( \mathbb{R}^{d_x \times d_e} \), \( \mathbb{R}^{d_e \times d_f} \), respectively. And \( b^{(1)}_f, b^{(2)}_f \) are the bias terms, \( \sigma \) is the activation function, here, we use ReLU function.

After obtaining the visual representations of the item collection \( \tilde{P} = \{x_1^f, x_2^f, ..., x_k^f\} \), we can generate the item embedding matrix \( M \in \mathbb{R}^{k \times d_f} \), where \( k \) is the maximum length of the item collections and \( d_f \) is the latent dimension of item visual feature. If the item collection length is shorter than \( k \), we repeatedly use a ‘padding’ item until the length is \( k \). Next, an attention mechanism is applied to model the item compositional coherence in multiple visual spaces. Inspired by the multi-heads attention method [Vaswani et al., 2017], in this paper, we use a stacked attention model to capture the global visual coherence. More specifically, we transform the item visual features into various visual spaces and perform a linear projection as \( x^*_s = x_i W_s + b_s \), where \( x^*_s \) is the representation vector of \( i \)-th item in \( s \)-th visual space, \( W_s \) is the projection matrix of shape \( \mathbb{R}^{d_f \times d_s} \), and \( b_s \) is the bias term. Then we can model the global coherence between two items in the target collection as:

\[
\alpha_{s_{i,j}} = \frac{W^q_s x^*_i \cdot (W^k_s x^*_j)^T}{\sqrt{d_s}}, \hat{\alpha}_{s_{i,j}} = \frac{\exp(\alpha_{s_{i,j}})}{\sum_{j=1}^{k} \exp(\alpha_{s_{i,j}})},
\]

where \( \alpha_{s_{i,j}} \) is the visual coherence score of item \( p_i \) and \( p_j \), weight matrices \( W^q_s, W^k_s \) are of shape \( \mathbb{R}^{d_s \times d_s} \). Here, we use \( \sqrt{d_s} \) to constraint the values of \( \alpha_{s_{i,j}} \). Next, the coherence scores are normalized by a softmax function. Then we can generate the item representation \( v^i_s \) incorporating the global coherence of other items in the \( s \)-th visual space as:

\[
v^i_s = \sum_{j=1}^{k} \hat{\alpha}_{s_{i,j}} \otimes x^*_j,
\]

where the \( \otimes \) represents the element-wise product. For each visual space, we can generate each item representation including the compositional information. Then, we can fuse all the visual space to generate the global item visual representations, i.e., \( v_i = Concatenate(v^1_i, v^2_i, ..., v^s_i) \). Therefore, we model global coherence from multiple visual spaces within an attention-based block and item visual representation \( M_a = [v_1, v_2, ..., v_k] \) is generated. Then, the global coherence learning process can be defined as \( f_a(P) \). However, the global coherence learning process is a linear operation and the
interactions between different latent dimensions are largely unexploited. With this in mind, we develop a feedforward network to endow the model with the nonlinear operation:

\[ f_n(M) = \sigma(M_nW_n^{(1)} + b_n^{(1)})W_n^{(2)} + b_n^{(2)}, \]

where the weight matrices \( W_n^{(1)}, W_n^{(2)} \) are of shape \( \mathbb{R}^{d_f \times d_f} \) and \( b_n^{(1)}, b_n^{(2)} \) are the bias terms. After that, we generated the attention output from a multi-heads attention block. For modeling the sophisticated compositional coherence of global features, we stack multiple multi-heads attention blocks to build a deeper network architecture. In order to prevent overfitting and unstable training process, we perform the Batch Normalization (BN) layer [Ioffe and Szegedy, 2015] after the output of each attention block. More formally, the output of \( b \)-th block can be defined as:

\[ M_a^{(b)} = f_a(h^{b-1}) \odot h^{b-1}, \]
\[ h^{(b)} = BN(W_mf_n(M_a^{(b)})) + b_{bn}, \]

where \( \odot \) is the element plus, \( f_a, f_n \) are the attention function and feedforward function, respectively. After the final attention block, we can generate global representations of the seed collection \( \mathcal{S}(P) = h^{(k)} \), where \( h^{(k)} \) is the output of the last multi-heads attention block.

3.2 Focal Coherence Learning

As mentioned above, for learning the compositional coherence from the item content, the content semantic attributes are also important to understanding item characteristics. To that end, we propose a novel Focal Coherence Learning (FCL) to model the compositional coherence from semantic-focal contents, i.e., color-focal contents, texture-focal contents and hybrid-focal contents. Firstly, we develop a Semantic-Focal Content Extraction (SFCE) based on the region-based segmentation methods as regions can yield much richer visual information than pixels [Uijlings et al., 2013]. Inspired by Selective Search [Uijlings et al., 2013], we firstly use a grouping algorithm to merge the semantic regions, which are based on the color or textual similarity computing. Specifically, we first generate the initial semantic regions by a segmentation algorithm [Felzenszwalb and Huttenlocher, 2004]. Then we develop the Selective Search algorithm to group regions which are similar in color, texture and both of them, respectively. The similarity can be measured as \( \text{Sim}(r_i, r_j) = \sum_{k=1}^{n} \min(C_i^k, C_j^k) \), where the \( C_i \) is the measurement of characteristic feature histogram, i.e., the color histogram and texture histogram, \( n \) is the feature dimensionality. More specifically, for each region, we extract three-dimensional color histograms. Each color channel is discretized into 25 bins, and \( n = 75 \) for the color feature in total. And we obtain \( n = 240 \) dimension texture histogram by computed SIFT-like [Liu et al., 2010] measurements on 8 different orientations of each color channel, for each orientation for each color using 10 bins. Both color histogram and texture histograms are normalized by \( L_1 \) norm.

After computing the semantic similarity of regions \( i \) and \( k \), we can group the similar regions and the target histogram \( C_i \) can be efficiently propagated by:

\[ C_t = \frac{\text{size}(r_i) \cdot C_i + \text{size}(r_j) \cdot C_j}{\text{size}(r_i) + \text{size}(r_j)}. \]

The size of target regions can be simply computed \( \text{size}(r_i) = \text{size}(r_i) + \text{size}(r_j) \). Different from the Selective Search [Uijlings et al., 2013], we do not combine the color and texture similarity. Because we want to deeply explore the compositional coherence of different semantic contents. CANN extracts the semantic-focal contents \( R_c, R_t, R_h \) respectively based on color similarity, texture similarity and hybrid similarity, and for each semantic-focal content we choose three content regions. With a pre-trained Inception V3 [Szegedy et al., 2016], we can generate the feature vectors \( F = [y_c, y_t, y_h] \) of all the semantic-focal contents. Next, we can learn the focal coherence among the content regions. Actually, considering the computational complexity of the attention mechanism for all the regions in all the items, similar with [Ma et al., 2019], we propose a hierarchical attention module to model both the semantic-specific and cross-semantic dependency in a distinguishable way. As shown in Figure 2, the input semantic-focal features \( V \in \mathbb{R}^{k \times |P| \times |R| \times d_f} \) can be reshaped as input matrices \( V_c \in \mathbb{R}^{k \times |P| \times d_f} \) in the semantic-specific space and \( V_S \in \mathbb{R}^{k \times |P| \times d_f} \) in the cross-semantic space. For each seed item collection, \( k \) is the seed item numbers, \( |P| \) is the number of semantic-focal features, \( |R| \) is the number of content regions of each semantic-focal feature, and \( d_f \) is the dimension of each region representation. Afterwards, we can compute our focal coherence by hierarchical multi-heads attention as:

\[ V' = AV = \hat{A}_C V_S = \hat{A}_C \hat{A}_S V, \]

where \( \hat{A}_C \) and \( \hat{A}_S \) are attention matrices of semantic-specific and cross-semantic. Similar to GBL, the attention matrices are computed by Eq.2. More specifically, we firstly model the compositional coherence of different semantic-focal regions. Then, we learn the attention map of different items. For aligning the attention matrices to the semantic-focal features, we reshape the attention matrices as \( \hat{A}_C = A_C \cdot I_C \) and \( \hat{A}_S = A_S \cdot I_S \), where \( I_C, I_S \) are the identity matrices. To that end, the semantic-focal representation of the item seed collection \( P \) can be formulated as \( \mathcal{F}(V) = V' \).

3.3 Optimization Strategy

So far, from GCL and FCL modules, we can generate the global and semantic-focal representations, respectively. Next, after a fully connected layer, we can obtain representations of the prediction items. Following [Han et al., 2017], we append a softmax layer on the \( \hat{x} \) to calculate the probability of complementary items conditioned on the target seed collections:

\[ \text{Pr}(\hat{x}|P) = \frac{\exp(\hat{x} \cdot x_i)}{\sum_{x_i \in \mathcal{N}} \exp(\hat{x} \cdot x_i)}, \]

where \( \mathcal{N} \) contains all the items from the seed collections. This can make model learn compositional coherence by means of considering a diverse collection of candidates. For optimizing our CANN, we propose a compositional training
Algorithm 1 Compositional Optimization Strategy

\textbf{Input:} Initialization model \( f(P, C; \theta) \); The length of the seed collection \( k \); The complementary item database \( S \); The number of epochs \( T \); The size of batch \( m \)

\textbf{Parameter:} Model parameter \( \theta \)

\begin{algorithmic}[1]
\For {\( i = 1, 2, 3, ..., T \)}
\State Random sample \( m \) seed collections \( O \in S \)
\State Initial input mini-batch \( \text{Input} \) as \( \emptyset \)
\For {\( O_i \) in \( \text{Batch} \)}
\State Random choose an item \( p \) from the collection \( O_i \)
\State Generate the seed collection \( P_i \leftarrow \text{Mask}(O_i, p) \)
\If {\( |P_i| < k \)}
\State Add an padding to the left of \( P_i \) until \( |P_i| = k \)
\EndIf
\EndFor
\State Generate the input mini-batch \( \text{Input} \leftarrow \text{Input} \cup P_i \)
\For {\( \text{Input} \)}
\State Build the training candidates \( \mathcal{N} \leftarrow \forall x \in \text{Input} \)
\State Update the model \( \theta \leftarrow \text{SGD}(f(\text{Input}, C; \theta), \theta) \)
\EndFor
\EndFor
\end{algorithmic}

strategy to simulate users’ decision-making process on complementary items as Algorithm 1. In the training stage, we use the \( \text{Mask}(O_i, p) \) operation to delete prediction items \( p \) from the seed collections \( O \). Moreover, we take all items in the mini-batch as the candidate collection \( \mathcal{N} \). But in the inference stage, we conduct the candidate collection as the \( x_c \). Actually, for some small datasets, we can use all the items as the candidates, but this is not practical for large datasets because of the high dimensional item representations.

With the compositional optimization strategy, CANN can be trained end-to-end. During training, we minimize the following objective function:

\[
L(P, \mathcal{N}; \theta) = -\frac{1}{|\mathcal{N}|} \sum_{i=1}^{|\mathcal{N}|} \log \Pr(\hat{x} | P),
\]

where \( \theta \) denotes the model parameters. Similar with [Han et al., 2017], we add a visual-semantic embedding as a regularization. We use Stochastic Gradient Decent (SGD) [Robbins and Monro, 1951] with mini-batch to update them through Algorithm 1.

4 Experiments

In this section, we first introduce the experimental settings and compared methods. Then, we compare the performance of CANN against the compared approaches on the complementary recommendation task. Then, we make an ablation study on the focal coherence learning process. At last, we conduct a case study to visualize the compositional coherence of our proposed CANN.

4.1 Experimental Setups

\textbf{Dataset.} We evaluate our proposed method on a real-world dataset, i.e., Polyvore dataset [Han et al., 2017; Vasileva et al., 2018]. It provides the fashion outfits, which are created and uploaded by experienced fashion designers. Indeed, outfit matching is a natural scenario for the content-based complementary recommendations, because the clothes in an outfit are complementary to each other. We use the provided datasets [Han et al., 2017; Vasileva et al., 2018], which contain 90,634 outfits. For the reliability of experimental results, we make the necessary specific processing as follows. First, we merge the datasets and remove the noise samples that exclude the seed sets of more than 8 items. It is because a fashion outfit hardly contains more than 8 items in real-world scenarios. Then, we split the dataset into 59,212 outfits with 221,711 fashion items for training, 3,000 outfits for validation and 10,218 outfits for testing. Next, we conduct a testing set (i.e., FITB_Random) as fill-in-the-blank tasks [Han et al., 2017], which we randomly choose candidates as the negative samples from the whole testing set. Meanwhile, in order to further make our testing more difficult and similar to users’ decision-making process, we conduct a more specific testing set, i.e., FITB_Category. In this set, we remove the easily identifiable negative samples and replace these samples with other items which have the same category of the ground-truth.

\textbf{Evaluation Metrics.} We evaluate our model and all compared methods by two evaluation metrics, i.e., the Accuracy (ACC) [Han et al., 2017; Vasileva et al., 2018] and Mean Reciprocal Rank (MRR) [Song et al., 2017; Jin et al., 2019].

\textbf{Implementation Details.} We adopt the GoogleNet InceptionV3 model [Szegedy et al., 2016] which was pretrained on ImageNet [Deng et al., 2009] to transform global visual contents (images) and semantic-focal visual contents (regions) to feature vectors. For fair comparisons, we set all the image embedding size of \( d_f = 512 \) unless otherwise noted. The number of visual space is set to \( S = 4 \) and for each visual space, we set \( d_s = d_f/S = 128 \) and \( b^* = 4 \) unless otherwise noted. Our model is trained with an initial learning rate of 0.2 and is decayed by a factor of 2 every 2 epochs. The batch size is set to 9, seed collection length \( k \) is set to 8. We stop the training process when the loss on the validation set stabilizes. Our model and all the compared methods are developed and trained on a Linux server with two 2.20 GHz Intel Xeon E5-2650 v4 CPUs and four TITAN Xp GPUs. The datasets and source codes are available in our project pages \(^1\).

4.2 Compared Approaches

To demonstrate the effectiveness of CANN, we compare it with the following alternative methods:

- \textbf{SetRNN} [Li et al., 2017]. This method treats the outfit data as a set and develops RNN model to generate the complementary scores of target item sets.
- \textbf{SiameseNet} [Veit et al., 2015]. This model utilizes a Siamese CNN to project two items into a latent space to estimate their similarity. We use an L2 norm to normalize the item visual embedding before calculating the Siamese loss and set the margin parameter to 0.8.
- \textbf{VSE} [Han et al., 2017]. Visual-Semantic Embedding (VSE) method learns a multimodal item embedding based on images and texts. The resulting embeddings are used to measure the item recommendation scores.
- \textbf{Bi-LSTM} [Han et al., 2017]. This method builds a bidirectional LSTM to predict the complementary item condi-

\(^1\)https://data.bdaa.pro/BDAA_Fashion/index.html
tioned on previously seen items in both directions. We use the full model by jointly learning the multimodal inputs.

- **CSN-Best** [Vasileva et al., 2018]. This method learns a visual content embedding that respects item type, and jointly learns notions of item similarity and compatibility in an end-to-end model. We use the full components with a general embedding size of 512 dimensions in our experiments.

- **NGNN** [Cui et al., 2019]. This method learns item complementary relations by the node-wise graph neural networks. This is the state-of-the-art method in fashion compatibility learning and clothing complementary recommendations.

Besides, for verifying the effectiveness of components, we construct three variant implements based on CANN.

- **CANN-G**. This is a variant of our model that only uses the Global Coherence Learning (GCL) module to generate complementary recommendations.

- **CANN-F**. This is a specific implementation that only uses Focal Coherence Learning (FCL) module.

- **CANN**. This is the model with all proposed components.

### 4.3 Recommendation Performances

We compare CANN with the other methods on the content-based complementary recommendations. The number of candidates is set to 4 for both FITB_Random and FITB_Category, which is the same as previous work [Han et al., 2017; Vasileva et al., 2018; Cui et al., 2019].

The results of all methods on both testing sets are shown in Table 1. We can make the following observations: 1) CANN outperforms all the compared methods in both datasets, which indicates the superiority of the proposed model for content-based complementary recommendations. 2) SetRNN and VSE perform the worst on both datasets. That indicates the complementary method is not a trivial problem which can be handled straightforward by simple methods. 3) SiameseNet and CSN-Best are similar methods which model the item pairwise compatibility. These models work better than SetRNN and VSE model, but worse than Bi-LSTM and CANN. This may be due to the fact that SiameseNet and CSN-Best aim to learn the similarity between two fashion items but ignore the compositional process. 4) CANN-F performs worse than CANN-G, that because CANN-F only considers the semantic focal contents, which may make some information ignored. Moreover, CANN-G outperforms other methods except for CANN. That enables us to safely draw the conclusion that it is advisable to model the compositional coherence of items on both global and semantic-focal contents.

### 4.4 Ablation Study on Focal Coherence

To further assess the robustness of the model and necessity of the semantic-focal coherence, we set different semantic contents in FCL module and evaluate the performances as the number of candidate collections increases. We set the FCL module with only color-focal, texture-focal and hybrid-focal contents and compare these models with all our proposed model CANN which contains all the semantic-focal contents.

The results are shown in the Figure 3, where the horizontal axis indicates the number of candidate collections. The

<table>
<thead>
<tr>
<th>Approaches</th>
<th>FITB_Random</th>
<th>FITB_Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>MRR</td>
</tr>
<tr>
<td>SetRNN</td>
<td>29.6%</td>
<td>48.1%</td>
</tr>
<tr>
<td>SiameseNet</td>
<td>52.2%</td>
<td>71.6%</td>
</tr>
<tr>
<td>VSE</td>
<td>29.2%</td>
<td>49.1%</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>83.6%</td>
<td>91.1%</td>
</tr>
<tr>
<td>CSN-Best</td>
<td>58.9%</td>
<td>76.1%</td>
</tr>
<tr>
<td>NGNN</td>
<td>87.3%</td>
<td>93.2%</td>
</tr>
<tr>
<td>CANN-G</td>
<td>88.8%</td>
<td>94.1%</td>
</tr>
<tr>
<td>CANN-F</td>
<td>71.9%</td>
<td>84.1%</td>
</tr>
<tr>
<td>CANN</td>
<td><strong>90.7%</strong></td>
<td><strong>95.1%</strong></td>
</tr>
</tbody>
</table>

Color, Texture, Hybrid and ALL represent that CANN uses only color-focal, texture-focal, hybrid-focal and all contents, respectively. From the Figure 3, we can get the following observations: 1) CANN with all the semantic-focal contents has outperformed others, which clearly demonstrate the effectiveness of all components in our proposed CANN. 2) The CANN with hybrid-focal coherence learning performs better than color-focal and texture-focal methods. Meanwhile, the color-focal model outperforms the texture-focal model. These observations imply that the color is an important factor for content-based complementary item recommendations. 3) CANN with all semantic-focal contents outperforms other single semantic-focal models on FITB_Category with a larger margin than on FITB_Random. These observations imply that semantic-focal contents can help the model to better understand the item compositional relationships and generate the best-matched complementary item suggestions.

### 4.5 Visualization of the Attention Mechanism

To further illustrate the learning and expression of compositional visual coherence in our model, we visualize the intermediate results of the coherence score $\alpha_{i,j}$ in Eq. 2. Figure 4 illustrates the item compositional relations in an example. For the better visualization, we select one semantic region generated by our model for each semantic-focal content. The color in Figure 4 changes from light to dark while the value of coherence score increases.

From Figure 4, it is worth noting that the coherence scores between the t-shirt and shorts are higher than others in all coherence spaces. Meanwhile, the scores between shoes and bracelet are also quite high. That is intuitive that the black t-shirt and dark blue shorts are similar in color. The shoes and bracelet are similar in leopard print style. These observations imply that our proposed CANN can provide a good way to capture the visual coherence for the complementary items from both global and semantic-focal views.

### 5 Conclusion

In this paper, we proposed a novel framework, the Content Attentive Neural Network (CANN), to address the problem of content-based complementary recommendations. For generating complementary item recommendations, the global and
semantic contents (such as color collocations, texture compatibilities) are indispensable parts to understand the comprehensive compositional relationship among items. Along this line, we provided a focused study on the compositional coherence in item visual contents. More specifically, we first proposed a Global Coherence Learning (GCL) module based on multi-heads attention to model the global compositional coherence. Then, we generated the content semantic-focal representations and designed a hierarchical attention module, i.e., Focal Coherence Learning (FCL), to learn the focal coherence from different semantic-focal contents. Next, for simulating users’ decision-making process on complementary items, we optimized the CANN in a novel compositional optimization strategy. Finally, we conducted extensive experiments on a real-world dataset and the experimental results clearly demonstrated the effectiveness of CANN compared with several state-of-the-art methods.

In the future, we would like to consider multi-modal information, such as item descriptions and categories for the deep exploration of the item complementary relationships. Moreover, we are also willing to investigate the domain knowledge about aesthetics assessment and make the recommender systems more explainable.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (Grants No. 61922073, 61672483, U1605251). Q. Liu acknowledges the support of the Youth Innovation Promotion Association of CAS (No. 2014299) and the USTC-JD joint lab. Bo Wu thanks the support of JD AI Research. We special thanks to all the first-line healthcare providers, physicians and nurses that are fighting the war of COVID-19 against time.

References


