Learning to Match on Graph for Fashion Compatibility Modeling

Xun Yang,[1*] **Xiaoyu Du,**[1,2*] **Meng Wang**[3]
[1]School of Computing, National University of Singapore
[2]School of Information and Software Engineering, University of Electronic Science and Technology
[3]Department of Computer Science, Hefei University of Technology
xunyang@nus.edu.sg, duxy.me@gmail.com, wangmeng@hfut.edu.cn

Abstract

Understanding the mix-and-match relationships between items receives increasing attention in the fashion industry. Existing methods have primarily learned visual compatibility from dyadic co-occurrence or co-purchase information of items to model the item-item matching interaction. Despite effectiveness, rich extra-connectivities between compatible items, e.g., user-item interactions and item-item substitutable relationships, which characterize the structural properties of items, have been largely ignored. This paper presents a graph-based fashion matching framework named Deep Relational Embedding Propagation (DREP), aiming to inject the extra-connectivities between items into the pairwise compatibility modeling. Specifically, we first build a multi-relational *item-item-user* graph which encodes diverse item-item and user-item relationships. Then we compute structured representations of items by an attentive relational embedding propagation rule that performs messages propagation along edges of the relational graph. This leads to expressive modeling of higher-order connectivity between items and also better representation of fashion items. Finally, we predict pairwise compatibility based on a compatibility metric learning module. Extensive experiments show that DREP can significantly improve the performance of state-of-the-art methods.

1 Introduction

Fashion has been an integral part of our everyday life. It is about not only what people wear, but also a mirror of people's attitude toward life, reflections of culture, arts, and even economics. It is a rapidly growing industry and has motivated various research topics in the fashion domain, such as recommendation (Yu et al. 2018; Zhang et al. 2017), search (Liu et al. 2016), and dialogue systems (Liao et al. 2018), etc.

In this paper, we focus on a newly-emerged topic of *Mix-and-match*-based fashion recommendation (Han et al. 2017; Vasileva et al. 2018; Song et al. 2017), for which the goal is to predict the matching score between fashion items from different categories. For example, when a user views/buys an item (e.g., a red floral maxi dress), the system matches it with the compatible fashion items from a complementary category (e.g., high-heel sandals). Most people, especially the young women/girls spend much time every day to think about the question of "What to wear" or "how to match the clothes for a good outfit". If there is an application that could help people to choose the suitable clothes, it would be very popular. That is the importance of this fashion matching task. The keys to solving the fashion matching problem are 1) how to represent fashion items, and 2) how to effectively model the item-item compatibility relationship.

Mainstream methods have primarily leveraged the visual appearance and side information of fashion items to learn visual compatibility in a pairwise learning manner (Veit et al. 2015; McAuley et al. 2015; He, Packer, and McAuley 2016; Chen and He 2018; Song et al. 2017; Yang et al. 2019a). A common assumption behind them is that a pair of compatible items should stay close with each other in a latent space. Then, the matching problem is solved under a similarity learning paradigm: first collect a corpus of matched/unmatched item pairs, and then train a parameterized similarity function that enforces matched pairs have a higher similarity score than unmatched pairs.

Despite promising progress, existing solutions mainly exploit dyadic co-occurrence (Song et al. 2017; Han et al. 2017) or co-purchase (Veit et al. 2015; He, Packer, and McAuley 2016) information of fashion items to model the item-item matching interaction. They forgo utilizing rich extra-connectivity information between compatible items, such as historical user-item interactions (e.g., rating) and item-item substitutable relationships (e.g. also-viewed information), thus being insufficient to capture the rich yet complicated matching patterns. We argue that the extra-connectivity information affiliated with items, which characterizes the structural properties of items in a real-world e-commerce environment, should be carefully taken into account to enhance the compatibility relationship modeling and item representation learning. For example, if users frequently co-purchase two fashion items from two complementary categories (e.g., dresses and sandals), the two items may have strong compatibility. If users view a fashion item and finally buy another one, the two substitutable items may be compatible to the same fashion items. Such extra-connectivities can effectively complement to the compatibility relationships.

Recent efforts have tried to alleviate the above-mentioned limitations by refining pairwise compatibility with category-

*Co-first author. These authors contributed equally to this work.

category complementary relationships (Yang et al. 2019b; Vasileva et al. 2018) or manually-designed clothing matching rules (Song et al. 2018). However, the category-category complementary relationships in (Yang et al. 2019b) only use coarse-grained categories to bridge two items from complementary categories, which have limited extra-connectivities. The matching rules in (Song et al. 2018), functioning as a kind of expert information, is hard to be well-defined by data scientists without strong domain knowledge. In summary, the rich extra-connectivity information, e.g., user rating behaviors, affiliated with items has not been fully explored for fashion matching.

To fill the research gap, this paper aims to inject the extra-connectivity information into the item-item compatibility relationship modeling in a graph-based fashion matching framework, named Deep Relational Embedding Propagation (DREP). Specifically, we first build a multi-relational *item-item-user* graph which encodes diverse item-item and user-item relationships. Then, we learn vector representations of items by an attentive relational embedding propagation rule that performs messages aggregation and propagation along edges of the multi-relational graph with an attention network. The introduced attention network enables our DREP attend the most informative neighbors under different relation types. This leads to expressive modeling of extra and higher-order connectivities between items and also better representation of fashion items. We further present a compatibility metric learning module for predicting compatibility, which can effectively capture intra-modality and inter-modality correlation based on the structured embeddings of items with multimodal descriptions.

Our contributions are summarized as follows.

- We present a graph-based fashion matching framework that injects the extra-connectivities of items into pairwise compatibility modeling by attentive relational embedding propagation.
- We propose a compatibility metric learning module to predict the pairwise item-item compatibility by capturing inter-modality and intra-modality feature correlations.
- We justify the effectiveness of the proposed DREP on the large-scale Amazon dataset with rich item-item relationships and user behaviors. Extensive experiments demonstrate the effectiveness of DREP.

2 Related Work

Fashion Compatibility Learning. Existing works can be mainly classified into two groups: one is outfit creation (Han et al. 2017; Hsiao and Grauman 2018) aiming to automatically compose fashion outfits , and the other one is modeling item-item compatibility (Chen and He 2018; Song et al. 2018; He, Packer, and McAuley 2016; Song et al. 2017; Vasileva et al. 2018; McAuley et al. 2015). Most existing methods in the second group cast fashion matching as a metric learning problem by assuming that a pair of matched items should be *close* to each other in a latent space. Earlier works model the pairwise compatibility with data-independent interaction functions, e.g., inner-product (Song et al. 2017), or Euclidean distance (McAuley et al. 2015; Chen and He 2018), which are improved by data-dependent interaction function, such as probabilistic mixtures of non-metric embeddings (He, Packer, and McAuley 2016), and category-aware conditional similarity (Yang et al. 2019b; Vasileva et al. 2018).

Our proposed DREP contributes a new solution for this task that injects the extra-connectivities (e.g., user rating behavior or item-item co-viewed information) of fashion items into item representations by designing an attentive embedding propagation architecture. Our work is also related to deep metric learning methods (Yang, Zhou, and Wang 2018; Yang, Wang, and Tao 2018), where we present a compatibility metric learning method to model the intra-modality and inter-modality feature correlations.

Graph Neural Networks. Graph neural networks (GNNs) have gained increasing attention in recent years (Kipf and Welling 2017; Zitnik, Agrawal, and Leskovec 2018; Wang et al. 2019; Cao et al. 2019a). It has become a power tool to model graph structured data by higher-order messaging passing on graph. One of the basic GNNs architectures is presented in (Kipf and Welling 2017), which has been extended to model relational data in (Schlichtkrull et al. 2018), and learn the weights of neighbors with an attention mechanism in single-relational graph (Veličković et al. 2018).

Our proposed DREP introduce GNNs into fashion compatibility learning over a large scale structured data with different relation types. Different with current GNNs architectures, we design an attentive embedding propagation layer for multi-relational data, which can effectively modulate the contributions of neighbors under different relation types in the messaging-passing procedure. Our proposed compatibility metric learning module can also enhance the node representation learning in an end-to-end manner, which would facilitate structured data modeling in other domains.

Multimedia Recommendation. Our work is also related to multimedia recommendation methods (Xu et al. 2018; Yu et al. 2018; Chen et al. 2017; 2019; Hou et al. 2019) which leverage visual information to enhance user-item interaction modeling. In (Chen et al. 2019), a multi-modal attention neural network is designed to generate visual explanation for explainable fashion recommendation. In (Hou et al. 2019), a semantic extraction network and Fine-grained Preferences Attention module are designed to project users and items into this fine-grained interpretable semantic space. In this work, we only focus on item-item compatibility relationship modeling for across-category fashion recommendation. However, we encode the user-item interaction behavior and item-item co-viewed information in *Amazon* dataset into a multi-relational graph for item embedding aggregation and propagation. We found that extra-connectivities are helpful to uncover unseen item-item compatibility patterns. It is also the first time to exploit user behaviors to model item-item compatibility relationship in a structured-data modeling framework.

3 Problem Formulation

In this paper, we formulate a graph-based fashion compatibility learning task. Let $\mathcal{G}$ denote an undirected item-item-user graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of a set of vertices $\mathcal{V}$ and

a set of edges $\mathcal{E} \subseteq \{(v_i, v_j, r_{ij}) | v_i, v_j \in \mathcal{V}, r_{ij} \in \mathcal{R}\}$, where $\mathcal{R}$ corresponds to two types of pairwise relationships: item-item relationships, and user-item relationships. The vertices $\mathcal{V}$ include two types of nodes: items $\{x_i\}_i^{|\mathcal{X}|} \in \mathcal{X}$ and users $\{u_i\}_i^{|\mathcal{U}|} \in \mathcal{U}$. The goal of our task is to build a predictive model that estimates the compatibility score between x_i and x_j from different categories: $\hat{y}_{ij} = f(x_i, x_j)$ where f denotes the predictive model, and $\hat{y}_{ij}$ denotes the predicted compatibility score of a pair of items. The compatibility relationship is defined by binary labels $\mathcal{Y} = \{y_{ij}\}$ between items from different categories, $y_{ij} = 1$ if x_i is compatible to x_j, otherwise 0. This paper aims to explore extra-connectivity information for fashion compatibility learning in a graph-based framework, which is formally defined as:

- **Inputs:** A corpus of fashion items $\mathcal{X}$ with pairwise compatibility relationships $\mathcal{Y}$, and a multi-relational graph $\mathcal{G}$ encoding extra-connectivity information between items: $\{\mathcal{X}, \mathcal{Y}, \mathcal{G}\}$. Each item x_i is represented by one or more feature vectors $\{\mathbf{x}_i^p\}_{m=1}^P$ extracted from multiple descriptions, e.g., images, textual descriptions, and so on.
- **Outputs:** A pairwise ranking function for each pair of items (x_i, x_j), i.e., $f : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which maps a pair of items to a compatibility score by jointly considering the compatibility relationship and extra-connectivities.

In this task, we expect two important characteristics of fashion items would be effectively leveraged for enhancing fashion compatibility modeling: (1) structural properties characterized by rich extra connectivities, (2) multi-modalities of fashion items, e.g, visual appearance, textural description, ID, etc.

4 Our Proposed Approach: DREP

This work develops a *Deep Relational Embedding Propagation* framework for fashion compatibility learning that mainly consists of two key modules:

- An embedding propagation architecture that derives low-dimensional item embeddings via an attentive relational embedding propagation rule.
- A compatibility metric learning layer that models the pairwise compatibility relationship by taking intra-modality and inter-modality correlation into modeling.

We describe how to extract original feature vectors of items as the input of embedding propagation layer in the latter part of section 4. The training process is performed in an end-to-end manner. During testing, given a pair of items, we first extract their original feature vectors and then feed them into the embedding propagation module for deriving structured item embeddings, and finally compute the compatibility score based on a learned metric.

Embedding Propagation on Graph

The basic idea behind DREP is to inject extra-connectivity information (e.g., user behaviors, item-item co-viewed information) of fashion items into the compatibility relationships modeling and item representation learning. We aim to design a message-passing mechanism (Gilmer et al. 2017) to transform and propagate item/user information on the graph $\mathcal{G}$. For this purpose, motivated by recent progress on graph neural network (Gilmer et al. 2017; Kipf and Welling 2017), we introduce a multi-layer embedding propagation architecture in DREP that can transform and propagate node information on the graph structure across different types of relations.

Layer-wise Embedding Propagation. The input to the k-th layer of DREP is a set of node features, $\mathbf{E}^k = \{\mathbf{e}_i^k \in \mathbb{R}^{d_k}\}_{i=1}^N$, where $N = |\mathcal{I}| + |\mathcal{U}|$ is the number of nodes (e.g., items and users) in $\mathcal{G}$, and d_k is the dimension of node features. The layer updates the node features as $\{\mathbf{e}_i^{k+1}\}_{i=1}^N$ by feature transformation and propagation. A basic form of layer-wise embedding propagation on multi-relational graph is presented in RGCN (Schlichtkrull et al. 2018), which takes the following rule:

$$\mathbf{e}_i^{k+1} = \sigma\Big(\sum_{r\in\mathcal{R}}\Big(\frac{1}{\sqrt{|\mathcal{N}_i^r||\mathcal{N}_j^r|}}\sum_{j\in\{i,\mathcal{N}_i^r\}}\mathbf{W}_r^k\mathbf{e}_j^k\Big)\Big), \quad (1)$$

where $\mathcal{N}_i^r$ denotes the set of neighbor indexes of node i under relation r, $\mathbf{W}_r^k \in \mathbb{R}^{d_k \times d_{k+1}}$ denotes a relation-specific linear transformation matrix, and σ denotes a non-linear element-wise activation function, such as rectified linear unit. Eq. (1) is used for updating all node embeddings on the graph $\mathcal{G}$ with the following propagation rule:

$$\mathbf{E}^{k+1} = \sigma\Big(\sum_{r\in\mathcal{R}}\tilde{\mathbf{D}}_r^{-\frac{1}{2}}\tilde{\mathbf{A}}_r\mathbf{D}_r^{-\frac{1}{2}}\mathbf{E}^k\mathbf{W}_r^k\Big) \quad (2)$$

where $\tilde{\mathbf{A}}_r = \mathbf{A}_r + \mathbf{I}$ is the adjacency matrix of the undirected graph $\mathcal{G}$ under relation r with added self-connections, $\mathbf{I}$ is the identity matrix, and $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$. Eq. (1) performs feature transformation and aggregation over the neighborhood structure of ego node, which can be seen as a kind of *message-passing* from the neighborhood of node i to node i. This leads to explicit modeling of second-order connectivities and information propagation between items (users). Specifically, for an item node in $\mathcal{G}$, it may gather signals from substitutable items: $x_i \leftarrow x_{j(i)}$ and also users that have rated it: $x_i \leftarrow u_{j(i)}$, which is a critical step to exploit the structural properties of items. However, the main weakness of such propagation rule is that it re-weights each neighbor of node i by a constant $1/\sqrt{|\mathcal{N}_i^r||\mathcal{N}_j^r|}$, which only depends on the structure of graph, being insufficient to discriminate the contributions of different neighbors under different types of relations.

Attentive Relational Embedding Propagation. To alleviate the limitedness of Eq. (1), we introduce an attentive relational propagation mechanism into DREP, which is motivated by the graph attention network (Veličković et al. 2018). Formally, the attentive relational embedding propagation is defined as following:

$$\mathbf{e}_i^{k+1} = \sigma\Big(\sum_{r\in\mathcal{R}}a^r\sum_{j\in\{i,\mathcal{N}_i^r\}}a_{ij}^r\mathbf{W}_r^k\mathbf{e}_j^k\Big), \quad (3)$$

where a_{ij}^r denotes the attention score of the neighbor j of ego node i under relation type r , which is computed by a

softmax operation

$$a_{ij}^r = \frac{\exp\left(\sigma'\left(\mathbf{h}_k^T\left(\mathbf{W}_r^k\mathbf{e}_i^k||\mathbf{W}_r^k\mathbf{e}_j^k\right)\right)\right)}{\sum_{j'\in\{i,\mathcal{N}_i^r\}}\exp\left(\sigma'\left(\mathbf{h}_k^T\left(\mathbf{W}_r^k\mathbf{e}_i^k||\mathbf{W}_r^k\mathbf{e}_{j'}^k\right)\right)\right)}, \tag{4}$$

where the attention mechanism is implemented by a single-layer feed forward neural network parameterized by a weight vector $\mathbf{h}_k \in \mathbb{R}^{2d_k\times 1}$ at the layer k of DREP. a^r denotes the weight of relation $r \in \mathcal{R}$, $\sum_{r\in\mathcal{R}} a^r = 1$ ($a^r \geq 0$). σ' is a non-linear activation function, and $(\cdot||\cdot)$ denotes the concatenation operation. By such a layer-wise attentive embedding propagation rule in Eq. (3), our DREP can effectively update the representation of nodes (items and users) based on the local area topological structure of the graph $\mathcal{G}$. All the node embeddings are updated using the following rule:

$$\mathbf{E}^{k+1} = \sigma\Big(\sum_{r\in\mathcal{R}} a^r\mathbf{A}_r'\mathbf{E}^k\mathbf{W}_r^k\Big), \tag{5}$$

where $\mathbf{A}_r' \in \mathbb{R}^{N\times N}$ denotes the attention matrix under relation type r, whose element at (i, j) is a_{ij}^r.

Higher-oder Embedding Propagation: By stacking K propagation layers defined in Eq. (3), a deeper model can be built for exploring higher-order messaging passing along the edges of the graph under different types of relations. This facilitates the expressive modeling of higher-order connectivities between items, such as $x_i \leftarrow u_{j(i)} \leftarrow i_{k(j)}$ or $x_i \leftarrow x_{j(i)} \leftarrow x_{k(j)}$. After a series of non-linear embedding transformation and aggregation operations, we obtain the structured node embeddings from the output of the last embedding propagation layer of DREP: $\{\mathbf{e}_1^K, \cdots, \mathbf{e}_i^K, \cdots, \mathbf{e}_N^K\}$. Note that among the structured node embeddings set, we only keep the item embeddings $\{\mathbf{x}_i^K \in \mathbb{R}^{d_K}\}_{i=1}^{|\mathcal{X}|}$ which have been enriched by the higher-order connectivities encoded in $\mathcal{G}$, since we focus on predicting item-item compatibility score. The final structured embedding of each item is denoted by

$$\mathbf{x}_i^* = \left(\mathbf{x}_i^0||\mathbf{x}_i^K\right), \tag{6}$$

where $\mathbf{x}_i^* \in \mathbb{R}^d$ is the final vector representation of item x_i, $d = d_0 + d_K$, and $(\cdot||\cdot)$ denotes the concatenation operation. In Eq. (6), both the original feature vectors and the output of embedding propagation module are preserved, which aims to avoid the degradation of the representation ability of item embeddings after multiple steps of embedding propagations.

Compatibility Metric Learning

After obtaining the structured embeddings of items by embedding propagation and aggregation, the next problem is how to predict pairwise compatibility score. Given a pair of items x_i and x_j with embeddings $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$, a common solution is modeling compatibility via a data-independent distance function

$$f(x_i, x_j) \propto -d(\mathbf{x}_i^*, \mathbf{x}_j^*) \tag{7}$$

where $d(\mathbf{x}_i^*, \mathbf{x}_j^*)$ is usually implemented by a squared Euclidean distance $\|\mathbf{x}_i^* - \mathbf{x}_j^*\|^2$ (Veit et al. 2015) or vector inner-product (Song et al. 2017). However, the data-independent distance function ignores complex statistical characteristic of embedding $\mathbf{x}_i^*$ from Eq. (6) and treats each dimension of vector $\mathbf{x}_i^*$ equally, being insufficient to model complex compatibility patterns. Recent work (He et al. 2018) on collaborative filtering has shown that data-dependent interaction function is better than data-independent function on modeling user-item interaction. Inspired by a recent work (He et al. 2018) on collaborative filtering, we design a data-dependent compatibility function with the following Mahalanobis-like distance

$$d_{\mathbf{M}}(x_i, x_j) = \|\mathbf{x}_i^* - \mathbf{x}_j^*\|_{\mathbf{M}}^2 \tag{8}$$

where $\mathbf{M} \in \mathbb{R}^{d\times d} \succeq 0$ is a positive semidefinite (PSD) matrix whose eigenvalues are nonnegative, which aims to leverage feature interaction for compatibility computing. *To ensure* $\mathbf{M}$ *to be PSD is critical in optimizing a Mahalanobis-like distance*. In this work, we consider the metric $\mathbf{M}$ as a diagonal matrix (Xing et al. 2003).

$$\mathbf{M} = \text{diag}\{m_1, \cdots, m_i, \cdots, m_d\}, \tag{9}$$

where m_i is the i-th main diagonal element of $\mathbf{M}$, defined as $m_i = d\eta_i^2/\sum_{j=1}^d \eta_j^2 \geq 0$, which ensures that m_i is smooth and nonnegative. By learning a diagonal metric $\mathbf{M}$ in Eq. (9), we expect that different dimensions of item embeddings are given different weights. Then, the goal is casted into optimizing a vector $\eta = (\eta_1, \cdots, \eta_d)$.

Multimodal Compatibility: An item x_i , usually, has multiple vector representations $\{\mathbf{x}_{i(p)}^0\}_{p=1}^P$ from P modalities, such as image, text, etc. We perform embedding propagation in DREP for each modality separately and obtain multiple structured embeddings $\{\mathbf{x}_{i(p)}^*\}_{p=1}^P$ for each item. For exploiting multimodal complementation, we extend Eq. (8) to

$$d_{\mathbf{M}}(x_i, x_j) = \frac{1}{P^2}\sum_{(p,q)} \|\mathbf{x}_i^{*p} - \mathbf{x}_j^{*q}\|_{\mathbf{M}_{(p,q)}}^2, \tag{10}$$

where $\mathbf{M}_{(p,q)} \succeq 0$ is a PSD metric that measures the compatibility across modality p and q. It is also defined as a diagonal matrix via Eq. (9). Eq. (10) provides an effective solution to model the compatibility relationship by learning a set of diagonal metrics, which explores not only intra-modality correlation and but also inter-modality correlation. Note that $\mathbf{M}_{(p,q)} = \mathbf{M}_{(q,p)}$.

Remarks: The reason of designing such a Compatibility Metric Learning module in this fashion compatibility framework is that the Euclidean distance or cosine similarity cannot work very well to capture the feature-level compatibility relationship with the concatenated multi-modality representation as input, due to the significant difference on the statistical distributions of multi-modalities. To address this issue, we design such a trainable and easy-to-implement multimodal fusion metric, which has shown good performance in the experiments.

Margin-based Ranking Criterion

Given a compatible item pair $(x_i, x_j) \in \mathcal{P}$, we randomly sample a set of negative items which do not have compatibility relationships with x_i or x_j in training set $\mathcal{P}$. In this

way, we generate a large set $\mathcal{T}$ of triplets for training:

$$\mathcal{T} = \{(x_i, x_j, x_l)|(x_i, x_l) \notin \mathcal{P} \cup (x_j, x_l) \notin \mathcal{P}\}. \quad (11)$$

To jointly optimize item structured embeddings and distance metrics $\mathbf{M}$, we minimize a margin-based ranking criterion over the training set

$$\mathcal{L} = \frac{1}{|\mathcal{T}|} \sum_{(i,j,l)\in\mathcal{T}} [d_{\mathbf{M}}(x_i, x_j) - d_{\mathbf{M}}(x_i, x_l) + \gamma]_+, \quad (12)$$

where $[\cdot]_+$ denotes hinge loss, and $\gamma > 0$ is a margin parameter, and $|\mathcal{T}|$ denotes the total number of training triples. The optimization goal is to ensure that the distance between a pair of compatible items is smaller than that between incompatible (randomly sampled) items by a margin.

Implementation Details

In this work, we mainly use two types of widely used features ($P = 2$) as original vector representations of items: one is visual features extracted from images, and one is textual features extracted from textual descriptions.

Item Visual features: We adopt a pretrained VGG network to extract visual features of items. Given an image of item x_i, the output of pretrained VGG is $\mathbf{x}_i^{cnn} \in \mathbb{R}^{d^{cnn}}$ ($d^{cnn} = 4096$). Then we apply a one-layer feed forward network $g_v(\cdot)$, parametrized by weight matrix $\mathbf{W}_v \in \mathbb{R}^{d_0 \times d^{cnn}}$ to transform the output of VGG into a d_0-dimensional vector $\mathbf{x}_{i(v)}^0$ as the input of DREP: $\mathbf{x}_{i(v))}^0 = \mathbf{W}_v \mathbf{x}_i^{cnn}$. Note that we just use the pretrained CNN features to represent the visual modality of each item to validate the effectiveness of our proposed approach. Actually, this visual module can be extended to more fancy one. It is easy to design a spatial attention neural network (Chen et al. 2019) to capture the region-wise visual features. When two fashion items have a high compatibility score, we can visualize the attention map of each item to give a visual explanation that reveals which part of this item make the main contribution to this match case. Another potential solution is to extract the regions of interest in each item image. Then, each item can be represented as a set of region features. We can also exploit the bilinear pooling layer (He and Chua 2017) to model the second-order interaction of different visual regions as a more expressive visual embedding.

Item Textual features: We adopt the pretrained *Glove* word embeddings (Pennington, Socher, and Manning 2014) to extract 300-d feature vectors of words in a sentence, and use average-pooling to aggregate word vectors as a sentence vector $\mathbf{x}_i^{glove} \in \mathbb{R}^{300}$ for each item. We also apply a one-layer feed forward network $g_t(\cdot)$, parametrized by weight matrix $\mathbf{W}_t \in \mathbb{R}^{d_0 \times d^{glove}}$ to transform the pretrained sentence vector into a d_0-dimensional dense vector: $\mathbf{x}_{i(t)}^0 = \mathbf{W}_t \mathbf{x}_i^{glove}$.

User ID embeddings: For the representation of users, we extract the ID embedding by an embedding looking up operation following the widely-used solution in collaborative filtering methods (He et al. 2018). The extracted ID embeddings of users are fed into DREP for embedding propagation.

Table 1: Statistics of the datasets.

Datasets	Amazon-Men	Amazon-Women
#Items	73,737	110,928
#Users	95,782	134,968
#Compatibility Rel.	367,230	467,623
#Substitute Rel.	352,529	517,423
#User-Item Interactions	253,468	377,693

5 Experiments

In this section, we conduct extensive experiments to justify the effectiveness of our proposed DREP on compatibility learning. 1) RQ1: Can DREP achieve competitive performance by exploiting the extra-connectivities? 2) RQ2: Can DREP effectively model the compatibility relationship? 3) RQ3: How do hyper-parameters effect the performance of DREP?

Dataset

In this work, we employ the widely-used *Amazon* (Men and Women) (Veit et al. 2015) dataset to justify the effectiveness of DREP on modeling compatibility relationship. Currently, only the *Amazon* dataset provides rich item-item relationships and user-item relationships simultaneously. Following the setting of (Veit et al. 2015), the pairwise compatibility relationship is defined as the *bought-together* and *also-bought* relationships between different categories. Two side relationships (extra-connectivities) provided in Amazon are used to build DREP: one is item-item substitute relationship (i.e., *also-viewed* information in Amazon), and the other is user-item interactions (i.e., *user rating* information in Amazon). Note that the *bought-together*, *also-bought*, and *also-viewed* relationships were not directly derived from customer behaviors, but from Amazon's recommendation algorithms. We evaluate our DREP on the sampled Amazon-Men and Amazon-Women datasets.

The statistics of our two datasets are shown in Table 1 (*Rel.* refers to *relationships*). We only use two side relationships to build the item-item-user graph. The compatibility relationship is used to sample the positive/negative training pairs for optimization under Eq. (12). We randomly sample 80% items for training, 10% items for validation, and 10% items for testing. Note that all query items in validation and testing sets do not have any compatibility relationships in the training set. In the validation and testing set, for each query item, we leave no more than 5 ground truths in candidate list.

Experimental Settings

Experimental Protocols: To evaluate top-K prediction of a ranked list, we use three metrics: 1) *Recall@K* that measures the fraction of compatible items, retrieved within the top K ranked list, out of all ground-truths. 2) *Hit@K* that measures the fraction of compatible items presented in the top K ranked list. 3) *NDCG@K* that accounts for the position of the hit by assigning higher scores to hits at top-K list. A higher Recall@K, Hit@K, or NDCG@K score denotes a better performance.

Table 2: Overall Performance Comparison with baseline methods in both the single modality setting (V, using visual image only) and multi-modality setting (V+T, using image and text). R@5, H@5, and N@5 refer to Recall@5, Hit@5, and NDCG@5. The higher score indicates a better performance.

Fea.	Method	Amazon-Men			Amazon-Women		
		R@5	H@5	N@5	R@5	H@5	N@5
V	SiaNet	0.449	0.516	0.378	0.421	0.471	0.343
	RGCN	0.507	0.582	0.412	0.515	0.581	0.428
	RGCN-ML	0.561	0.633	0.467	0.582	0.663	0.502
	DREP	**0.587**	**0.659**	**0.498**	**0.623**	**0.694**	**0.538**
V+T	SiaNet	0.576	0.651	0.495	0.547	0.610	0.465
	BPR-DAE	0.564	0.628	0.500	0.538	0.591	0.472
	RGCN	0.609	0.691	0.518	0.616	0.684	0.527
	RGCN-ML	0.719	0.795	0.646	0.713	0.784	0.642
	DREP	**0.735**	**0.805**	**0.668**	**0.728**	**0.793**	**0.663**

Baseline methods: We compare our proposed DREP with the following baseline methods to justify its effectiveness:

- **Siamese Nets** (Veit et al. 2015) (**SiaNet**). It measures visual compatibility using ℓ_2-normalized Euclidean distance.
- **BPR-DAE** (Song et al. 2017). This work models the pairwise compatibility as the inner-product of item embeddings using both images and textual descriptions as input.
- **RGCN** (Schlichtkrull et al. 2018). This work designs a relational graph neural network method for learning structured node embedding on a graph with different types of relations. It is implemented in the same framework with DREP. By default setting, RGCN uses Euclidean distance to model pairwise compatibility. Besides, to be fairly compared with our DREP, RGCN is also combined with the metric learning module (Eq. (8)), termed as **RGCN-ML**.

Among the listed methods, SiaNet and BPR-DAE do not rely on any extra-connectivity information, while RGCN can utilize multi-types of extra-connectivity information, which is a strong baseline for DREP. Except for BPR-DAE that is optimized with the Bayesian Personalized Ranking (BPR) objective (Rendle et al. 2009) using the visual and textual information, all the other methods are optimized with the margin ranking loss with a margin 0.5 and evaluated with both single modality setting (Visual only) and multi-modality setting (Visual and Textual). For the multi-modality fusion in RGCN (V+T) and BPR-DAE, we directly fuse the normalized low-dimensional embeddings $\mathbf{x}_i^0(v)$ and $\mathbf{x}_i^0(t)$ with an average pooling operation, resulting in better performance than vector concatenation or score-level summation.

Parameter settings: We implement DREP using Tensorflow. The number of embedding propagation layers in DREP is set to 2 by the performance on validation set. The embedding size of each layer is set to 64 for simplicity, resulting in 128-D item embeddings (as shown in Eq. (6)) from the output of DREP. We optimize all models with the Adagrad optimizer. The learning rate and regularization term are both fixed at 0.01 and 1e-5 by grid searching on validation set. Except for BPR-DAE, all the embedding vectors are normalized to unit one for stable learning. We report the performance of all the methods on testing set using the model with best performance on validation set.

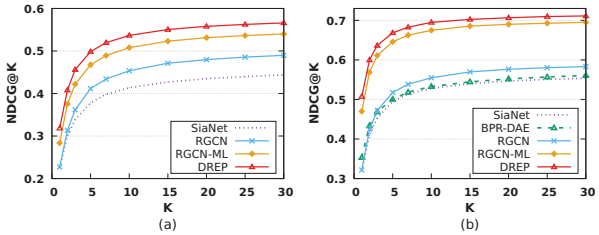

Figure 1: Top-K prediction comparison *w.r.t.* NDCG@K at different ranking positions on Amazon-Men. (a) Only visual modality is utilized, (b) multi-modality setting is adopted.

Overall Performance Comparison

Table 2 displays the performance comparison *w.r.t.* **R**ecall@K, **H**it@K, and **N**DCG@K (K=5) on *Amazon-Men* and *Amazon-Women*. Figure 1 reports top-K performance comparison *w.r.t.* NDCG@K at different ranking positions on Amazon-Men. We have the following findings:

- RGCN-ML, RGCN, and our DREP consistently outperform SiaNet and BPR-DAE which do not use extra-connectivities by large margins. It indicates the necessity of exploiting the rich extra-connectivity information for learning item representation with a graph neural network. That makes sense since rich extra-connectivities are helpful to recover unseen item-item compatibility interactions by embedding propagation on the item-item-user graph.
- RGCN-ML substantially surpasses RGCN. It is reasonable since our compatibility metric learning module can learn a diagonal metric that assigns different weights on different dimensions of item embeddings, which can effectively capture feature correlation for compatibility prediction. Especially, we ensure that each derived diagonal matrix satisfies the PSD condition, which makes the overall learning procedure more stable. The improvements are more significant in the multi-modality setting, benefiting from the intra-modality and inter-modality correlation in Eq. (10).
- Our DREP consistently achieves better performance than RGCN-ML, which verifies the effectiveness of our proposed attentive relational embedding propagation rule. DREP injects attention mechanism into the multi-relational embedding propagation which can modulate the contribution of different types of extra-connectivity information and also attend the informative neighbors for information passing along relation-specific edges.

Study of DREP

Analysis on Different Levels of Extra-connectivity: One of the main advantages of DREP is the injecting of extra-connectivity information into item-item compatibility modeling. We investigate the effectiveness of extra-connectivity

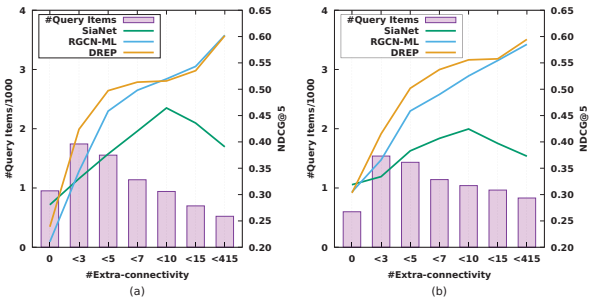

Figure 2: Comparison over seven subgroups of query items in testing set with different sparsity levels of extra-connectivities on Amazon-Men. Only visual modality is adopted. (a) denotes comparison *w.r.t.* item-item substitute relationship, (b) denotes comparison in a multi-relational setting (Both item-item and user-item extra connectivities are used).

information on different subgroups of query items in testing set with different sparsity levels. Specifically, we split the original query item set into seven subgroups based on the number of extra connectivity relationships per query item has. Figure 2 (a) and (b) show the comparison between DREP and RGCN-ML over seven subgroups of query items on Amazon-Men. The performances of SiaNet on different subgroups are reported as a baseline. We have the following observations:

- DREP and RGCN-ML slightly underperform SiaNet in the first subgroup since the query item has no connectivity in the graph, and performs consistently better in all the other subgroups. It indicates exploiting extra connectivity can greatly enhance the representation learning of items by capturing the structural characteristics of items.
- We also observe that DREP and RGCN-ML can yield remarkable improvement with only no more than 5 extra relationships. It is reasonable since items can interact with indirectly-connected neighbors on the graph by higher-order embedding propagation. The more extra relationships provided, the higher prediction score achieved. When the number of extra-relationships are large than 10, the performance of SiaNet drops fast. It reflects that only relying on visual appearance of items cannot provide a good prediction for very active items.

Convergence of DREP: Figure 3 shows the training loss curve and accuracy curve (*w.r.t.* NDCG@5) versus different number of epochs on validation set of Amazon-Men. We can observe that both RGCN-ML and DREP converge stably and fast within 200 epochs. Both methods can achieve significant performance improvement within top 100 epochs. Since we exploit the graph neural network module to encode the item-to-item and item-to-user structural relationships. When the number of nodes in such a multi-relational graph is very large, it will make the model very hard to optimize. A potential solution can be exploited to address this issue is that the whole graph can be partitioned into multiple sub-graphs for batch-wise training (Ying et al. 2018).

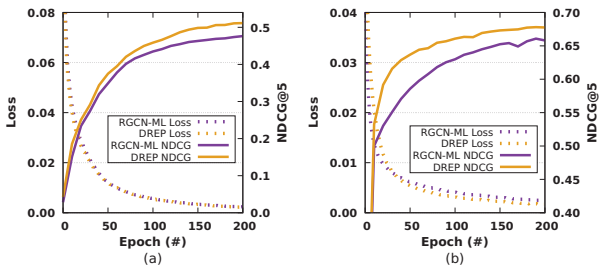

Figure 3: Loss curve on training set and accuracy curve (*w.r.t.* NDCG@5) on validation set of Amazon-Men. (a) Only visual modality is utilized, (b) multi-modality setting is adopted.

6 Conclusion

This paper developed a Deep Relational Embedding Propagation (DREP) framework for learning fashion compatibility, which aimed to inject extra-connectivities into the pairwise compatibility relationship modeling of fashion items. Specifically, we first build a item-item-user graph which encodes two types of extra-connectivity information, and user-item interactions. Then, we design a non-linear relational embedding propagation rule with an attention mechanism to modulate the contribution of each neighbor under different relation types, thus leading to expressive and structured item embeddings. Besides, we also design a compatibility metric learning module to better leverage the structural characteristics for compatibility modeling, instead of the data-independent compatibility function. Finally, by stacking multiple relational embedding propagation layers and the compatibility metric learning layer, we can effectively capture the higher-order connectivity information for the item-item compatibility prediction. Extensive experiments demonstrate that our proposed DREP can yield state-of-the-art fashion compatibility learning performance.

In the future, we would consider to encode more fashion domain knowledge (Ma et al. 2019) into the graph and also try to utilize the knowledge graph completion approach (Cao et al. 2019b) to handle new fashion items.

7 Acknowledgments

This research is part of NExT++ research and also supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61725203 and Grant 61732008. NExT++ research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

References

Cao, Y.; Liu, Z.; Li, C.; Li, J.; and Chua, T.-S. 2019a. Multi-channel graph neural network for entity alignment. In *ACL*, 1452–1461.

Cao, Y.; Wang, X.; He, X.; Hu, Z.; and Chua, T.-S. 2019b. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *WWW*, 151–161. ACM.

Chen, L., and He, Y. 2018. Dress fashionably: Learn fashion collocation with deep mixed-category metric learning. In *AAAI*, 2103–2110.

Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; and Chua, T.-S. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*, 335–344. ACM.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *ACM SIGIR*, 765–774. ACM.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *ICML*, 1263–1272. JMLR. org.

Han, X.; Wu, Z.; Jiang, Y.-G.; and Davis, L. S. 2017. Learning fashion compatibility with bidirectional lstms. In *ACM MM*, 1078–1086. ACM.

He, X., and Chua, T.-S. 2017. Neural factorization machines for sparse predictive analytics. In *ACM SIGIR*, 355–364. ACM.

He, X.; Du, X.; Wang, X.; Tian, F.; Tang, J.; and Chua, T.-S. 2018. Outer product-based neural collaborative filtering. In *IJCAI*, 2227–2233.

He, R.; Packer, C.; and McAuley, J. 2016. Learning compatibility across categories for heterogeneous item recommendation. In *ICDM*, 937–942. IEEE.

Hou, M.; Wu, L.; Chen, E.; Li, Z.; Zheng, V. W.; and Liu, Q. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. In *IJCAI*.

Hsiao, W.-L., and Grauman, K. 2018. Creating capsule wardrobes from fashion images. In *CVPR*, 7161–7170.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Liao, L.; Ma, Y.; He, X.; Hong, R.; and Chua, T.-S. 2018. Knowledge-aware multimodal dialogue systems. In *ACM MM*, 801–809. ACM.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 1096–1104.

Ma, Y.; Yang, X.; Liao, L.; Cao, Y.; and Chua, T.-S. 2019. Who, where, and what to wear?: Extracting fashion knowledge from social media. In *ACM MM*, 257–265. ACM.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, 43–52. ACM.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 452–461. AUAI Press.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *ESWC*, 593–607. Springer.

Song, X.; Feng, F.; Liu, J.; Li, Z.; Nie, L.; and Ma, J. 2017. Neurostylist: Neural compatibility modeling for clothing matching. In *ACM MM*, 753–761. ACM.

Song, X.; Feng, F.; Han, X.; Yang, X.; Liu, W.; and Nie, L. 2018. Neural compatibility modeling with attentive knowledge distillation. In *SIGIR*, 5–14.

Vasileva, M. I.; Plummer, B. A.; Dusad, K.; Rajpal, S.; Kumar, R.; and Forsyth, D. 2018. Learning type-aware embeddings for fashion compatibility. In *ECCV*, 390–405.

Veit, A.; Kovacs, B.; Bell, S.; McAuley, J.; Bala, K.; and Belongie, S. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 4642–4650. IEEE.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*. accepted as poster.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T.-S. 2019. Kgat: Knowledge graph attention network for recommendation. In *KDD*.

Xing, E. P.; Jordan, M. I.; Russell, S. J.; and Ng, A. Y. 2003. Distance metric learning with application to clustering with side-information. In *NeurIPS*, 521–528.

Xu, Q.; Shen, F.; Liu, L.; and Shen, H. T. 2018. Graphcar: Content-aware multimedia recommendation with graph autoencoder. In *ACM SIGIR*, 981–984.

Yang, X.; He, X.; Wang, X.; Ma, Y.; Feng, F.; Wang, M.; and Chua, T.-S. 2019a. Interpretable fashion matching with rich attributes. In *ACM SIGIR*, 775–784.

Yang, X.; Ma, Y.; Liao, L.; Wang, M.; and Chua, T.-S. 2019b. Transnfcm: Translation-based neural fashion compatibility modeling. In *AAAI*.

Yang, X.; Wang, M.; and Tao, D. 2018. Person reidentification with metric learning using privileged information. *TIP* 27(2):791–805.

Yang, X.; Zhou, P.; and Wang, M. 2018. Person reidentification via structural deep metric learning. *IEEE TNNLS* (99):1–12.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph convolutional neural networks for web-scale recommender systems. In *ACM SIGKDD*, 974–983. ACM.

Yu, W.; Zhang, H.; He, X.; Chen, X.; Xiong, L.; and Qin, Z. 2018. Aesthetic-based clothing recommendation. In *WWW*, 649–658.

Zhang, X.; Jia, J.; Gao, K.; Zhang, Y.; Zhang, D.; Li, J.; and Tian, Q. 2017. Trip outfits advisor: Location-oriented clothing recommendation. *TMM* 19(11):2533–2544.

Zitnik, M.; Agrawal, M.; and Leskovec, J. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34(13):i457–i466.