

# TransNFCM: Translation-Based Neural Fashion Compatibility Modeling

Xun Yang,<sup>1</sup> Yunshan Ma,<sup>1</sup> Lizi Liao,<sup>1</sup> Meng Wang,<sup>2</sup> Tat-Seng Chua<sup>1</sup>

<sup>1</sup>School of Computing, National University of Singapore, Singapore

<sup>2</sup>School of Computing and Information Engineering, Hefei University of Technology, China  
xunyang@nus.edu.sg, yunshan.ma@u.nus.edu, liaolizi.llz@gmail.com, eric.mengwang@gmail.com  
chuats@comp.nus.edu.sg

## Abstract

Identifying mix-and-match relationships between fashion items is an urgent task in a fashion e-commerce recommender system. It will significantly enhance user experience and satisfaction. However, due to the challenges of inferring the rich yet complicated set of compatibility patterns in a large e-commerce corpus of fashion items, this task is still underexplored. Inspired by the recent advances in multi-relational knowledge representation learning and deep neural networks, this paper proposes a novel Translation-based Neural Fashion Compatibility Modeling (TransNFCM) framework, which jointly optimizes fashion item embeddings and category-specific complementary relations in a unified space via an end-to-end learning manner. TransNFCM places items in a unified embedding space where a category-specific relation (category-comp-category) is modeled as a vector translation operating on the embeddings of compatible items from the corresponding categories. By this way, we not only capture the specific notion of compatibility conditioned on a specific pair of complementary categories, but also preserve the global notion of compatibility. We also design a deep fashion item encoder which exploits the complementary characteristic of visual and textual features to represent the fashion products. To the best of our knowledge, this is the first work that uses category-specific complementary relations to model the category-aware compatibility between items in a translation-based embedding space. Extensive experiments demonstrate the effectiveness of TransNFCM over the state-of-the-arts on two real-world datasets.

## Introduction

Recently, rising demands for fashion products have driven researchers in e-commerce to develop various techniques to effectively recommend fashion items. Existing techniques can be mainly categorized into two types: (1) *search*-based: when a user views a fashion item online, system suggests similar items which the user may also like, and (2) *mix-and-match*-based: when a user views a fashion item (e.g., blouse), the system suggests compatible items (e.g., pants) from a complementary category. The first one has already been used in most fashion e-commerce sites, by modeling the visual similarities or interaction relationships such as *also-viewed*. The second approach is more challenging and

still underexplored, since it needs to infer the *compatibility* relationships among fashion items that go beyond merely learning visual similarities. This paper mainly focuses on the second type, termed as cross-category fashion recommendation (CCFR). It has received increasing attention from multiple research fields, due to its potential ability to enhance user online shopping experience and satisfaction (Hsiao and Grauman 2018; Han et al. 2017; Vasileva et al. 2018; Veit et al. 2015).

The key to designing a CCFR model is 1) how to represent various fashion items, and 2) how to model their compatibility relationships based on their representations. Mainstream studies handle this task in an embedding learning strategy: one optimizes a feature mapping from original item space into a latent compatibility space, where compatible items are close together by requiring the pairwise Euclidean distance (inner-product) between embeddings of compatible fashion items to be much smaller (larger) than that of incompatible items (Veit et al. 2015; McAuley et al. 2015; Song et al. 2017; 2018; He, Packer, and McAuley 2016). Here, fashion items are represented as a latent vector (also termed as *embedding*) in a latent space, where cross-category compatibility is modeled with pairwise Euclidean distance or inner-product between the embeddings of fashion items.

Despite their promising performance, such similarity/metric learning-based compatibility modeling approaches usually suffer from the following limitations. (1) They just consider compatibility learning as a *single-relational* data modeling problem and use a fixed and data-independent function, i.e., Euclidean distance or inner-product, to model the notion of compatibility, which ignores the rich yet complicated patterns in fashion compatibility space. (2) They only utilize pairwise labels (i.e., compatible/incompatible) to optimize item embeddings, thus resulting in a context-unaware compatibility space which ignores the inter-category variation. In such a category-unaware space, incompatible items may be forced to close together in a *one-to-many* case due to the *improper similarity transitivity* (Vasileva et al. 2018). We argue that fashion compatibility is a multi-dimensional concept that varies from case to case. Fashion experts usually use different attributes of items to make a decision for different categories. Data-independent compatibility function usually results in sub-optimal performance.

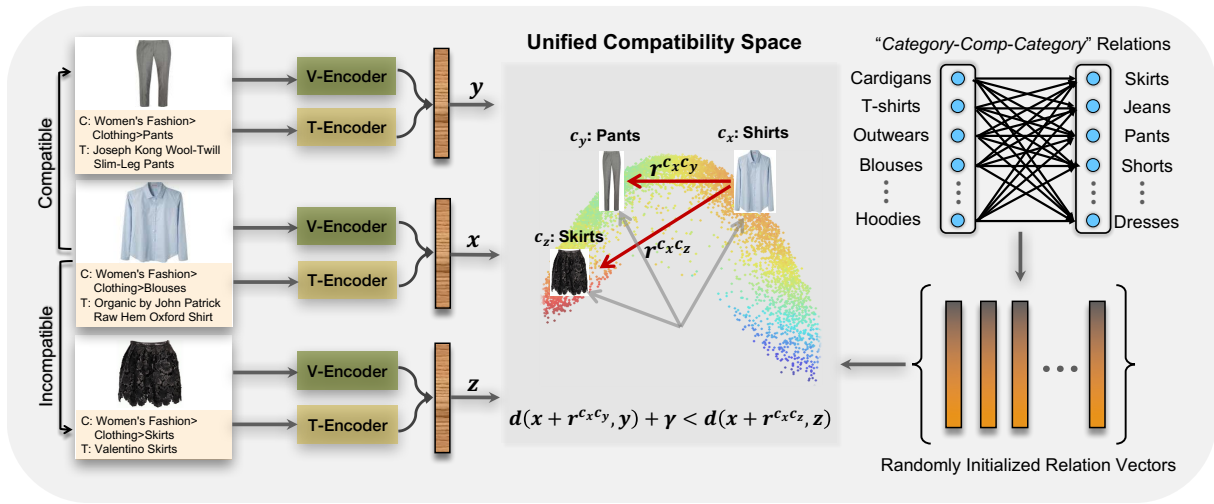


Figure 1: Overview of the proposed Translation-Based Neural Fashion Compatibility Modeling (TransNFCM) approach. It mainly consists of three parts: (1) Each item is first mapped into a latent space by a multimodal item encoder which consists of a pretrained deep CNN for visual modality and a text CNN for textual modality, (2) Category complementary relations (category-comp-category) are encoded into the latent space as vector translations operating on the embeddings of compatible items, and (3) Both item embeddings and relation vectors are jointly optimized by minimizing a margin-based ranking criterion.

To overcome these limitations, this paper proposes to design a data-dependent compatibility function which takes category labels of items into modeling. Specifically, we formulate compatibility learning as a *multi-relational* data modeling problem in a unified compatibility space dominated by a group of category complementary relations (category-comp-category) as shown in Figure 1. Here, each relation  $r$  corresponds to a complementary category pair  $(c_x, c_y)$  (e.g., T-shirts and Pants), connecting a head item  $x$  from  $c_x$  to a tail item  $y$  from  $c_y$ . Then, inspired by the highly-celebrated Translation Embedding (TransE) method (Bordes et al. 2013), we develop a Translation-Based Neural Fashion Compatibility Modeling (TransNFCM) approach by interpreting each relation as a *semantic translation* operating on the low-dimensional embeddings of fashion items. Given a pair of compatible items  $(x, y)$ , our basic learning objective is that the embedding of tail item  $x$  should be close to that of head item  $y$  plus the relation vector  $r^{c_x c_y}$ , i.e.,  $x + r^{c_x c_y} \approx y$ . Finally, we model the category-aware compatibility of  $(x, y)$  with a data-dependent distance function  $P((x, y) \in \mathcal{P}) \propto -d(x + r^{c_x c_y}, y)$ , which respects the category labels of the head item and tail item and yields a specific compatibility conditioned on  $r^{c_x c_y}$ . Besides, we design a neural item encoder to encode the visual and textual features of an item into a unified embedding vector that are jointly optimized with relation vectors in a latent space. By the proposed TransNFCM, we expect to learn a category-aware notion of compatibility which can better capture complicated compatibility patterns, e.g., *one-to-many*, in a single space.

Our contributions are summarized as follows.

- We present a novel end-to-end neural network architecture TransNFCM for fashion compatibility modeling.

This is the first work that uses category complementary relations to model category-respected compatibility between items in a translation-based embedding space.

- We exploit the multimodal complementary characteristic of fashion items with a neural item feature encoder.
- We conduct extensive experiments on two public datasets, which demonstrates the effectiveness of TransNFCM.

## Related Work

**Fashion Compatibility.** In recent years, substantial prior work has been devoted to model the human notion of fashion compatibility for fashion recommendation. Hu et al. (2015) proposed a personalized outfit recommendation method which models the user-items interaction based on tensor decomposition. A recurrent neural network method in (Han et al. 2017) models outfit composition as sequential process, implicitly learning compatibility via a transition function. Hsiao et al. (2018) proposed to create capsule wardrobes from fashion images. The above-mentioned approaches focus on outfit compatibility and do not explicitly model item-to-item compatibility.

Mcauley et al. (2015) and Veit et al. (2015) proposed to model the visual complementary relations between items by posing it as a metric learning task (Yang, Wang, and Tao 2018) that combines Siamese CNNs with co-purchase data in *Amazon* as supervision. Chen et al. (2018) proposed a triplet loss-based metric learning method (Yang, Zhou, and Wang 2018) for fashion collocation. Song et al. (2017; 2018) proposed to model compatibility as inner-product between items under Bayesian Personalized Ranking (BPR) framework (He and McAuley 2016) using co-occurrence relationships in *Polyvore* outfits data as supervision. Both visual and textual modalities are used to represent items with an inter-

modality consistency constraint in deep embedding learning manner. In (Song et al. 2018), domain knowledge rules are utilized for compatibility modeling in a teacher-to-student learning scheme (Hu et al. 2016). The main limitation is that they use a fixed and data-independent interaction function to model pairwise compatibility, which ignores the rich yet complicated compatibility patterns and then results in sub-optimal compatibility modeling performance. They will be hard to handle the *one-to-many* case in fashion domain, e.g., they may push incompatible items together improperly when they are compatible with the same target. (He, Packer, and McAuley 2016) proposed to alleviate this limitation by modeling the notion of compatibility with a weighted sum of pairwise distances in multiple latent metric spaces.

Different with the aforementioned methods, our work proposes to model category-aware compatibility by jointly optimizing category complementary (co-occurrence) relations and item embeddings in a unified latent space and interpreting the relations as vector *translations* among compatible items from the corresponding categories. The closet work to ours is (Vasileva et al. 2018) that jointly learns type-respected visual item embedding with type-specific sparse masks for modeling both pairwise similarity and category-specific compatibility, in which its category-specific mask projects item pairs into a type-specific subspace for computing compatibility. Our TransNFCM follows a similar motivation with (Vasileva et al. 2018), but designed in a different way. As analyzed in latter section, TransNFCM not only captures category-specific notion of compatibility but also preserves the global notion of compatibility, thus showing better performance than (Vasileva et al. 2018).

**Personalized Recommendation.** Our work is also related to personalized (user-item) recommendation methods (He et al. 2017; 2018b; 2018a; Tay, Anh Tuan, and Hui 2018) which focus on learning data-dependent interaction functions instead of using a fixed function and show state-of-the-art performance. Our work can also be considered as a content-based recommendation method (Chen et al. 2017; Yu et al. 2018) where the item features have richer semantics than just a user/item ID. In this paper, we mainly focus on cross-category recommendation in fashion domain. In the future, we will consider to utilize the *user* information (age, shape, weight, religion, etc) for personalized fashion recommendation.

**Knowledge Embedding.** Our method is inspired by recent advances in knowledge representation learning (Bordes et al. 2013; Xie et al. 2016; Xie, Liu, and Sun 2016; Lin et al. 2015; Nie et al. 2015), where the objective is to model multiple complex relations between pairs of various entities. One highly-celebrated technique TransE (Bordes et al. 2013) embeds relations of multi-relational data as translations operating on the low-dimensional embeddings of entities. TransE has been employed for visual relation modeling (Zhang et al. 2017) and recommendation (Tay et al. 2018), (He, Kang, and McAuley 2017). Our work is motivated by those findings. We treat each fashion item as an entity in knowledge graphs. Our key idea is to represent the category-aware compatibility relations between items as translations in a unified embedding space, as shown in Figure 1.

## Proposed Approach: TransNFCM

This paper aims to tackle the task of fashion compatibility learning, which needs to address two sub-problems:

- How to effectively represent fashion items that are usually described by multimodal data (e.g., video, image, title, description, etc)?
- How to effectively model the notion of category-aware compatibility between fashion items?

We address these two problems by developing an end-to-end deep joint embedding learning framework, termed as Translation-based Neural Fashion Compatibility Modeling. The overall framework is illustrated in Figure 1. Fashion items with multimodal descriptions are encoded as low-dimensional embeddings in a latent compatibility space via a multimodal item encoder. Then, the complicated compatibility relations are captured by a category-specific vector translation operation in a latent space.

### Multimodal Item Encoder

In the mainstream recommender systems (He et al. 2017), only the item ID is used to represent the item, resulting in a high-dimensional and very sparse feature input. While, content-based item representation (Chen et al. 2017) is more popular in the fashion domain (He and McAuley 2016; He, Packer, and McAuley 2016; Liao et al. 2018), since it can capture rich semantical features of fashion items. Generally, fashion items in e-commerce sites are described by multimodal data. This work aims to exploit the complementary characteristic of multimodal descriptions for robust item representation by designing a two-channel item encoder consisting of a visual encoder (**V-Encoder**), and a textual encoder (**T-Encoder**). We propose to simultaneously learn two nonlinear feature transformations: one is V-Encoder  $f_V(v_x)$  that maps an image  $v_x$  of item  $x$  into a visual feature space  $\mathcal{R}^d$  and the other is T-Encoder  $f_T(t_x)$  that transforms an textual description  $t_x$  of item  $x$  into a textual feature space  $\mathcal{R}^d$ . We implement both V-Encoder and T-Encoder using the popular deep convolutional network models. Both the outputs of V-Encoder and T-Encoder are  $\ell_2$ -normalized and concatenated as the final representation  $\mathcal{R}^{2d}$  of item  $x$  in a feature-level fusion manner.

### Fashion Compatibility Modeling

How to model the item-to-item compatibility is the key of this task. Let's first briefly recall the strategies in prior work. Given a pair of compatible items  $(x, y)$  from complementary categories and their feature vectors  $\mathbf{x} \in \mathcal{R}^D$ ,  $\mathbf{y} \in \mathcal{R}^D$ , the notion of compatibility is modeled as four ways:

- **Inner-product** has been used in previous works (Song et al. 2017; 2018; He and McAuley 2016) which model the compatibility as

$$P((x, y) \in \mathcal{P}) \propto \mathbf{x}^T \mathbf{y}, \quad (1)$$

where  $P((x, y) \in \mathcal{P})$  denotes the probability of  $x$  and  $y$  being compatible with each other.

- **Euclidean distance** is adopted in (McAuley et al. 2015; Veit et al. 2015; Chen and He 2018), which models the compatibility as  $P((x, y) \in \mathcal{P}) \propto -d(x, y)$ , where

$$d(x, y) = \|\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\mathbf{x}^T \mathbf{y}. \quad (2)$$

- **Probabilistic mixtures of multiple distances** is proposed in (He, Packer, and McAuley 2016) which models the compatibility with a weighted sum of distances  $d_k(x, y)$  in  $M$  metric subspaces, parameterized by  $M$  matrices  $\{\mathbf{E}_k\}_k^M$

$$\begin{aligned} d(x, y) &= \sum_k P(k|x, y) d_k(x, y) \\ &= \sum_k P(k|x, y) \|\mathbf{E}_0^T \mathbf{x} - \mathbf{E}_k^T \mathbf{y}\|_2^2, \end{aligned} \quad (3)$$

where  $P(k|x, y)$  denotes the probability of  $\mathbf{E}_k$  being used for pair  $(x, y)$ .

- **Conditional similarity** (Veit et al. 2017) is used in (Vasileva et al. 2018) to model the type-aware compatibility with the pairwise distance in a conditioned similarity subspace

$$d(x, y) = \|\mathbf{x} \odot \mathbf{w}^{c_x c_y} - \mathbf{y} \odot \mathbf{w}^{c_x c_y}\|_2^2, \quad (4)$$

where  $\odot$  denotes element-wise multiplication,  $c_i, c_j$  denote the types (categories) of item  $x$  and  $y$ , and  $\mathbf{w}^{c_i c_j}$  is a sparse vector, acting as a gating function that selects the relevant dimensions of the embedding for determining compatibility in the similarity subspace depending on the item type pair  $(c_i, c_j)$ .

**Remarks:** Both the first and second types of methods use a fixed and data-independent compatibility function (Eq. (1) or (2)). When embeddings of item are normalized to unity norm, the compatibility only depends on  $\mathbf{x}^T \mathbf{y}$ . It can only capture the *global* notion of compatibility in a shared embedding space, which ignores the rich yet complicated matching patterns in original item space. The third measurement (Eq. (3)) adopts a data-dependent function to model the *global* notion of compatibility as a probabilistic mixture of multiple local compatibility scores. However, it needs to optimize  $M+2$  projection matrices with limited constraints, which can easily get stuck in bad local optima. Eq. (4) models the conditioned compatibility by employing a mask operation to embed item pairs into type-specific subspaces for computing conditional similarity, which only captures the *specific* notion of compatibility.

**Translation-based Compatibility Modeling.** Most prior work just treats compatibility learning as a problem of single-relational data modeling towards learning a global notion of compatibility. However, fashion compatibility is a multi-dimensional concept that varies from case to case. For example, fashion items belong to various categories (e.g., dresses, coats, skirts, sandals, etc). Given a pair of item from category  $A$  and  $B$ , the visual/semantic attributes which fashion experts use for making a decision may be color, pattern, sleeve-length, material, etc. While, given an item pair from category  $C$  and  $D$ , the attributes they use would change accordingly. In this case, the category labels of items would influence the decision making of experts significantly.

To overcome the limitations of existing work, we propose to incorporate category complementary (i.e., co-occurrence) relationships into compatibility modeling, which formulates this task as a problem of multi-relational data modeling, towards learning a category-aware compatibility notion. For simplicity, we term category complementary relations as category-comp-category, such as dresses-comp-boots, which are illustrated in Figure 2. The next question then is how to explicitly represent such relations. Inspired by the highly-celebrated TransE method (Bordes et al. 2013), we interpret the category-comp-category relations as a simple vector translation operating on the embeddings of compatible items from the corresponding categories. Given a pair of compatible items  $(x, y) \in \mathcal{P}$  with embedding vectors  $(\mathbf{x}, \mathbf{y})$ , and the corresponding category-comp-category relation vector  $\mathbf{r}^{c_x c_y}$ , we assume that the embedding of tail item  $y$  should be close to that of head item  $x$  plus the relation  $\mathbf{r}^{c_x c_y}$ , i.e.,  $\mathbf{x} + \mathbf{r}^{c_x c_y} \approx \mathbf{y}$ . Then, our notion of compatibility is modeled with a data-dependent distance function conditioned on  $\mathbf{r}^{c_x c_y}$ , i.e.,  $P((x, y) \in \mathcal{P}) \propto -d(\mathbf{x} + \mathbf{r}^{c_x c_y}, \mathbf{y})$ , where

$$\begin{aligned} d(\mathbf{x} + \mathbf{r}^{c_x c_y}, \mathbf{y}) &= \|\mathbf{x} + \mathbf{r}^{c_x c_y} - \mathbf{y}\|_2^2 \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + \|\mathbf{r}^{c_x c_y}\|_2^2 \\ &\quad - 2\underbrace{\mathbf{x}^T \mathbf{y}}_{\text{global}} - 2\underbrace{(\mathbf{y} - \mathbf{x})^T \mathbf{r}^{c_x c_y}}_{\text{category-specific}}, \end{aligned} \quad (5)$$

During the training stage, the embeddings of items and relations are jointly optimized in an end-to-end manner. During the recommendation stage, candidates that have smaller distance (computed using Eq. (5)) with the query item would be top-ranked and suggested to user.

**Remarks:** Note that, as shown in Eq. (5), our compatibility modeling function not only preserves the **global** notion of compatibility by  $\mathbf{x}^T \mathbf{y}$ , but also captures the **category-specific** compatibility  $(\mathbf{y} - \mathbf{x})^T \mathbf{r}^{c_x c_y}$ , which is significantly different with (Vasileva et al. 2018). Here, the relation  $\mathbf{r}^{c_x c_y}$  functions as a *mask* implicitly that can select category-aware features of items for computing compatibility in a subspace conditioned on the pairwise categories  $(c_x, c_y)$ .

**Margin-based Ranking Criterion.** Given a positive triplet  $(x, y, \mathbf{r}^{c_x c_y})$  consists of two items  $(x, y) \in \mathcal{P}$  from category  $c_x$  and  $c_y$ , we generate a set of negative (incompatible) triplets with either the head or tail item replaced by a random item (but not both at the same time). In this way, we can generate a large set  $\mathcal{T}$  of 5-tuples for training:

$$\begin{aligned} \mathcal{T} &= \{(x, y, \mathbf{r}^{c_x c_y}, y', \mathbf{r}^{c_x c_{y'}}) | (x, y') \notin \mathcal{P}\} \cup \\ &\quad \{(x, y, \mathbf{r}^{c_x c_y}, x', \mathbf{r}^{c_{x'} c_y}) | (x', y) \notin \mathcal{P}\} \end{aligned} \quad (6)$$

where  $x'(x)$  and  $y(y')$  are not compatible with each other but from complementary categories. To jointly optimize item embeddings and category-specific relation vectors, we minimize a margin-based ranking criterion over the training set

$$\mathcal{L} = \sum_{\mathcal{T}} [d(\mathbf{x} + \mathbf{r}^{c_x c_y}, \mathbf{y}) - d(\mathbf{x}' + \mathbf{r}^{c_{x'} c_y}, \mathbf{y}') + \gamma]_+ \quad (7)$$

where  $(x', y') \in \{(x, y') \cup (x', y)\}$ , and  $[\cdot]_+$  denotes hinge loss, and  $\gamma > 0$  is a margin parameter, and  $|\mathcal{T}|$  denotes the total number of 5-tuples in training set. The optimization goal is that distance between a pair of compatible items should be smaller than that between incompatible (less compatible) items by a margin.

**Implementation Details.** The V-Encoder is implemented by the pretrained AlexNet (Krizhevsky et al. 2012) which consists of 5 convolutional layers and 3 fully-connected (FC) layers. We drop the last FC layer and add a new FC layer as our visual embedding layer that transforms the 4096-D output of the 2nd FC layer into a  $d$ -dimensional embedding. The T-Encoder is implemented with the text CNN architecture (Kim 2014) consisting of a convolutional layer, a max-pooling layer, and a FC layer. We use four filter windows with sizes of  $\{2, 3, 4, 5\}$  with 50 feature maps each. The textual data are first preprocessed by filtering words appearing in very less items and with very less characters, and then we represent each word with the publicly-available 300-D *word2vec* vector. We use a new FC layer to replace the last layer to transform the output of max-pooling layer into a  $d$ -dimensional textual embedding ( $d = 128$ ). In the multimodal fusion setting, the outputs of the V-Encoder and T-Encoder are  $\ell_2$  normalized and then concatenated as the final item embedding  $\mathcal{R}^{2d}$ . Note that other multimodal feature fusion strategies, such as score-level fusion, can also be used in TransNFCM.

We optimize our proposed TransNFCM with the objective Eq. (7) by stochastic gradient descent (SGD) in minibatch training mode. All the visual/textual item embeddings are normalized to unit norm. For each pair of complementary categories (i.e., t-shirts and pants), we generate a category-aware relation vector, having the same dimension with item embeddings. The relation vectors are randomly initialized and only normalized to unit norm at the beginning of optimization. No other regularization or norm constraints are enforced on relation vectors. TransNFCM is implemented with Pytorch. In each epoch, we first shuffle all the 5-tuples in  $\mathcal{T}$  and get a mini-batch in a sequential way.

## Experiments

To comprehensively evaluate the effectiveness of our proposed TransNFCM method, we conduct experiments to answer the following research questions:

**RQ1:** Can our proposed TransNFCM outperform the state-of-the-art fashion compatibility learning methods?

**RQ2:** Is the multimodal embedding fusion strategy helpful for improving the learning performance?

**RQ3:** Why does TransNFCM work?

## Experimental Settings

**Datasets.** We conduct experiments on two public fashion compatibility datasets, both crawled from the fashion social commerce site *Polyvore* (www.polyvore.com), which enables fashion bloggers to share their fashion tips by creating and uploading outfit compositions. In the following two datasets, we both use the item pairs that are co-occurring in an outfit as supervision for training and also evaluation.

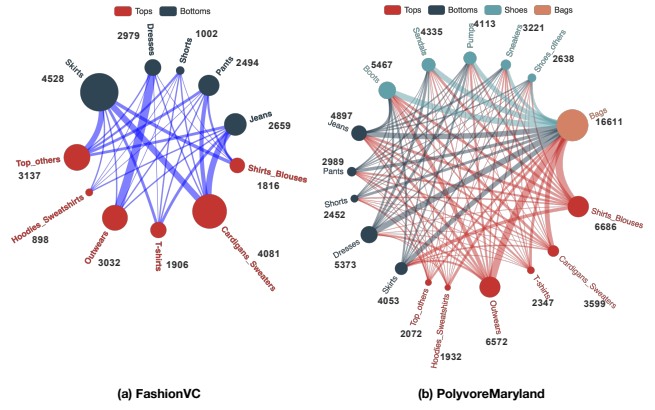


Figure 2: Illustration of category-comp-category relations in both datasets: (a) FashionVC, (b) PolyvoreMaryland. Each edge denotes a category-comp-category relation. The numbers of items in each category are also illustrated. For simplicity, *Dresses* is classified into the *Bottoms*. 30 category-comp-category relations are used in FashionVC and 101 relations are used in PolyvoreMaryland.

**FashionVC** (Song et al. 2017). It was collected for top-bottom recommendation, consisting of 14,871 tops and 13,663 bottoms, split into 80% for training, 10% for validation, and 10% for testing. Each item contains a product image and textual description (category and title).

**PolyvoreMaryland** (Han et al. 2017). It was released by (Han et al. 2017) for outfit compatibility modeling. Since in this paper we mainly study the modeling of item-item compatibility, not the whole outfit, we only keep four groups of items among the first 5 item in outfits: tops, bottoms, shoes, and bags. Other fashion accessories and jeweleries have been removed. Each item contains a product image and textual description (product title only). We extract and re-split the co-occurring item pairs randomly in the same setting with FashionVC: 80% for training, 10% for testing, and 10% for validation.

**Evaluation Protocols.** For each testing positive pair  $(h_i, t_{ig}) \in \mathcal{P}_t$ , we replace the tail item with  $N = 100$  negative items<sup>1</sup>  $\{t_{in}\}_{n=1}^N$  which do not co-occur with  $h_i$  in the same outfit but are from complementary categories with  $h_i$ . We adopt two popular metrics, *Area Under the ROC curve* (AUC) and *Hit Ratio* (HR), to evaluate the item-item recommendation based on the compatibility score. Both AUC and HR@K are widely-used in recommendation systems. AUC is defined as

$$\text{AUC} = \frac{1}{N|\mathcal{P}_t|} \sum_i \sum_n \delta(s(h_i, t_{ig}) > s(h_i, t_{in})) \quad (8)$$

where  $\delta(a)$  is an indicator function that returns 1 if the argument  $a$  is true, otherwise 0.  $s(h_i, t_{ig}) = P((h_i, t_{ig}) \in \mathcal{P}_t)$  denotes the compatibility score.  $|\mathcal{P}_t|$  denotes the total number of testing positive pairs. HR@K is a recall-based metric,

<sup>1</sup>Note that our setting is much more challenging than that of (Song et al. 2017; 2018) where only 3 negative candidates are sampled for each query during testing.

Table 1: Comparison on the FashionVC and PolyvoreMaryland datasets based on two metrics: AUC (%) and Hit@K (%),  $K \in \{5, 10, 20, 40\}$ ). A larger number indicates a better result. **V** and **T** denote **V**isual modality and **T**extual modality, respectively. **V+T** denotes the fusion of visual and textual modalities. 100 negative candidates are sampled for each query during testing. The best results are shown in boldface.

Features	Methods	FashionVC					PolyvoreMaryland				
		AUC	Hit@5	Hit@10	Hit@20	Hit@40	AUC	Hit@5	Hit@10	Hit@20	Hit@40
V	SiaNet	60.4	9.7	18.1	31.2	52.8	59.1	8.3	15.5	29.0	51.8
	Monomer	70.2	16.9	28.6	45.8	69.1	70.5	17.6	28.9	45.7	69.0
	CSN	71.6	16.7	28.4	46.7	70.8	70.2	17.3	28.4	45.1	68.4
	BPR	70.9	16.7	27.3	46.7	70.4	69.5	17.3	28.2	43.9	67.5
	TriNet	70.6	16.3	28.0	45.7	69.6	70.1	18.1	28.7	44.9	68.3
	<b>TransNFCM</b>	73.6	19.0	32.3	51.6	74.0	71.8	18.9	30.6	48.1	70.5
T	SiaNet	66.1	10.8	21.0	37.9	61.1	62.3	8.3	16.2	32.0	56.3
	Monomer	68.8	16.5	26.9	42.1	64.8	63.3	10.1	18.8	33.9	58.1
	CSN	67.5	11.2	22.4	41.2	64.1	63.2	8.8	17.0	32.5	57.4
	BPR	70.9	15.4	26.8	45.6	67.6	67.8	13.0	23.6	40.3	65.3
	TriNet	71.3	16.5	28.9	46.4	69.2	68.4	13.7	24.4	41.5	65.8
	<b>TransNFCM</b>	72.6	18.9	30.0	47.9	70.8	68.8	14.7	25.8	42.2	66.0
V+T	<b>TransNFCM</b>	76.9	23.3	38.1	57.1	77.9	74.7	21.7	34.4	52.7	75.3

denoting the proportion of the correct tail item  $t_{ig}$  ranked in the top  $K$ .  $N + 1$  candidates are provided for each head item  $h_i$ .

**Parameter Settings.** We employ the stochastic gradient descent for optimization with momentum factor as 0.9. We set the overall learning rate  $\eta = 0.001$ , and drop it to  $\eta = \eta/10$  every 10 epochs. The learning rate of the pretrained AlexNet in V-Encoder is set to  $\eta' = \eta/10$  for fine-tuning. The margin  $\gamma$  is set to 1 following the setting of (Bordes et al. 2013). We use 128 5-tuples in a minibatch. Both the dimensions of visual embedding vectors and textual embedding vectors are set to  $d = 128$ . Dropout is used in both visual and textual encoders.

**Comparison Methods.** We compare TransNFCM with the following methods that are all implemented in the same framework. The main difference lies in the compatibility modeling functions and loss functions.

- **Siamese Network** (SiaNet) (Veit et al. 2015): models compatibility with the squared Euclidean distance (Eq. (2)) and uses contrastive loss as its optimization criterion.
- **Triplet Network** (TriNet) (Wang et al. 2014; Chen and He 2018): models compatibility with Eq. (2) and use the margin-based ranking criterion with  $\ell_2$ -normalized item embeddings as input, in which the margin is set to 1. TriNet is our baseline which is category-unaware and uses data-independent function.
- **BPR** (He and McAuley 2016; Song et al. 2017): models compatibility as inner-product (Eq. (1)) and uses the soft-margin based objective loss. In this paper, we implemented it with  $\ell_2$ -normalized embeddings.
- **Monomer** (He, Packer, and McAuley 2016): models compatibility with a mixture of multiple local distances (Eq. (3)) and is also implemented using the same objective function with us.
- **CSN** (Vasileva et al. 2018): models pairwise compatibility as a conditional similarity (Eq. (3)) that respects item

types. We implement it following the setting of (Veit, Bellingie, and Karaletsos 2017).

## Experimental Evaluation

**Performance Comparison and Analysis.** Table 1 shows the performance comparison on two datasets based on AUC and Hit@K ( $K \in \{5, 10, 20, 40\}$ ). From Table 1, we have the following observations.

- TransNFCM achieves the best performance in most cases and obtains high improvements over the comparison methods, especially in the visual (V) modality setting. This justifies the effectiveness of TransNFCM that builds a data-dependent (i.e., category-aware) compatibility function (Eq. (5)) using a translation-based joint embedding learning paradigm. (**RQ1**)
- TransNFCM consistently outperforms the baseline TriNet on two datasets, especially using visual features. It demonstrates the necessity of encoding the pairwise category-labels into the embedding space for capturing a category-aware compatibility notion. Note that the improvement becomes less significant in the T setting on the PolyvoreMaryland dataset. It is mainly because that the textual descriptions of items in the PolyvoreMaryland dataset are very noisy and sparse, thus resulting in sub-optimal relation embeddings. (**RQ1**)
- Although both employing the pairwise category labels, TransNFCM significantly outperforms CSN in all. It is mainly because CSN only captures the conditioned compatibility in a subspace dominated by a pair of complementary categories and does not preserve the global notion of compatibility in original feature space, thus resulting in a limitation that its conditioned compatibility in one subspace cannot be compared with that in another subspace. While, benefiting from the TransE framework, our compatibility function (Eq. (5)) not only captures the



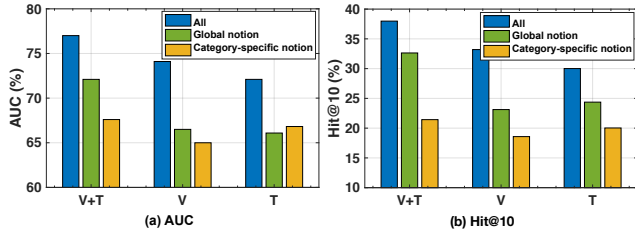


Figure 3: Effects of using different parts in Eq. (5) for compatibility computing on FashionVC. **Global notion** refers to using  $\mathbf{x}^T \mathbf{y}$ , **Category-specific notion** refers to using  $(\mathbf{y} - \mathbf{x})^T \mathbf{r}^{c_x c_y}$ , and **All** refers to using Eq. (5).

category-specific notion but also preserves the global notion, thus showing better performance. (RQ1)

- Table 1 also shows that TransNFCM can effectively mine the complementary characteristic of images and texts with the multimodal item encoder. The textual descriptions of items in the two datasets are very noisy and also sparse. But it indeed contributes some visually-invisible attributes such as *style* or *season* that can well complement visual features, thus bringing significant performance improvement. Note that BPR has been used for fashion compatibility modeling in a multimodal fusion manner in (Song et al. 2017). We have conducted a performance comparison with BPR in our multimodal fusion setting with 100 negative candidates: BPR achieves 75.4% AUC and 33.4% Hit@10 on the FashionVC dataset, and 73.8% AUC and 32.7% Hit@10 on the Polyvore dataset. TransNFCM yields remarkable improvements on BPR. Besides, we do not include a cross-modality consistency loss (Song et al. 2017) or visual-semantic loss (Vasileva et al. 2018) in TransNFCM, since we empirically found that those strategies would reduce the diversity of multimodal item representation and degrade the performance. (RQ2)

**Empirical analysis.** To better evaluate the effects of the global part and category-specific part in Eq. (5), we conduct an experiment (as shown in Figure 3) to investigate how these jointly-learned two parts performs separately in the testing stage. From Figure 3, we have the following observations.  $\mathbf{x}^T \mathbf{y}$  usually performs better than  $(\mathbf{y} - \mathbf{x})^T \mathbf{r}^{c_x c_y}$  in all except one, which demonstrates the importance of preserving such global notion of compatibility. The category-specific part complements the global part well, since it alleviates the *improper similarity transitivity* by pushing hard negative candidates farther away from the query. Only using  $(\mathbf{y} - \mathbf{x})^T \mathbf{r}^{c_x c_y}$  results in a bad performance, which is also validated by the performance of CSN in Table 1. In TransNFCM,  $\mathbf{r}^{c_x c_y}$  dominates a latent compatibility subspace where compatible items  $(x, y)$  from categories  $(c_x, c_y)$  are close to each other. So, we empirically conclude that it is the combination of both global notion and category-specific notion of compatibility that makes TransNFCM work. (RQ3)

**New setting: Target category is known.** We also compare TransNFCM with TriNet in a new evaluation setting: the tar-

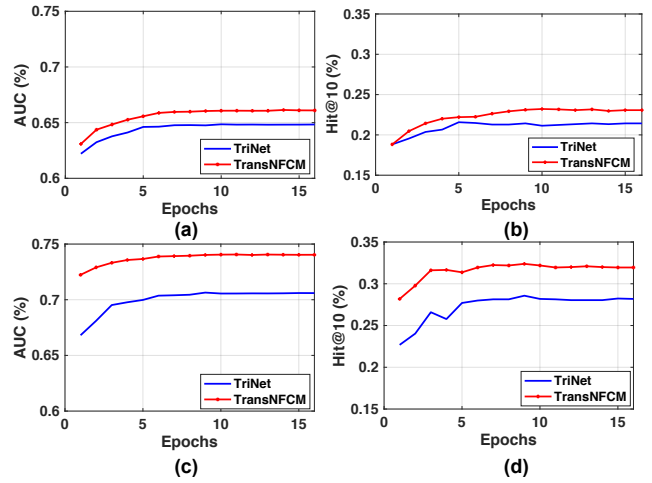


Figure 4: (a) (b) denote the comparison with TriNet when target category is known, (c) (d) denote the comparison with TriNet when target category is unknown. Experiments are conducted on FashionVC. Only visual modality is used.

get category is known, e.g., a user buys/clicks a blouse and wants system only recommend skirts that match the given blouse well. This new setting facilitates our baseline TriNet more, since all the negative candidates are sampled from the known target category. The *improper similarity transitivity* in TriNet can be alleviated. We observe from Figure 4 (a) and (b) that our TransNFCM can still outperform TriNet by nearly 2% AUC and Hit@10. When using the original setting (negative candidates are randomly sampled from complementary categories), shown in Figure 4 (c) and (d), the improvement becomes more obvious. Nearly 4% improvement is observed at both AUC and Hit@10.

## Conclusion

In this work, we developed a new neural network framework for fashion compatibility modeling, named TransNFCM. The main contribution is that TransNFCM casts fashion compatibility as a multi-relational data modeling problem and encodes the category-comp-category relations as vector translations operating on embeddings of compatible items in a latent space. TransNFCM not only captures the category-specific notion of compatibility but also preserves the global notion, which can effectively alleviate the *improper similarity transitivity* in metric learning based approaches. To the best of our knowledge, this is the first work that poses compatibility modeling into such a translation-based joint embedding learning framework. Although TransNFCM only utilizes category-level co-occurrence relations in this work, it can be directly extended to model fine-grained matching rules, composed of color, category, pattern, style, etc. We will explore more fine-grained relations in the fashion domain to further discover the potentials of TransNFCM in the future. Besides, we also introduce a multimodal item encoder which effectively exploits the complementary characteristic of different modalities. Extensive experiments on

two datasets have demonstrated the effectiveness of our method.

## Acknowledgments

We would like to thank all reviewers for their comments. NExT research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Funding Initiative. This work is also supported by the National Nature Science Foundation of China (NSFC) under grant 61732008 and 61725203.

## References

- Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.
- 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.
- 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, R., and McAuley, J. 2016. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *AAAI*, 144–150. AAAI Press.
- He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *WWW*, 173–182.
- He, X.; Du, X.; Wang, X.; Tian, F.; Tang, J.; and Chua, T.-S. 2018a. Outer product-based neural collaborative filtering. In *IJCAI*, 2227–2233.
- He, X.; He, Z.; Song, J.; Liu, Z.; Jiang, Y.-G.; and Chua, T.-S. 2018b. Nais: Neural attentive item similarity model for recommendation. *IEEE TKDE* 30(12):2354–2366.
- He, R.; Kang, W.-C.; and McAuley, J. 2017. Translation-based recommendation. In *RecSys*, 161–169. ACM.
- He, R.; Packer, C.; and McAuley, J. 2016. Learning compatibility across categories for heterogeneous item recommendation. In *ICDM*, 937–942. IEEE.
- Hsiao, W.-L., and Grauman, K. 2018. Creating capsule wardrobes from fashion images. In *CVPR*, 7161–7170.
- Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing deep neural networks with logic rules. In *ACL*, volume 1, 2410–2420.
- Hu, Y.; Yi, X.; and Davis, L. S. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *ACM MM*, 129–138. ACM.
- Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, 1746–1751.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.
- Liao, L.; Ma, Y.; He, X.; Hong, R.; and Chua, T.-S. 2018. Knowledge-aware multimodal dialogue systems. In *ACM MM*, 801–809. ACM.
- Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, volume 15, 2181–2187.
- McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, 43–52. ACM.
- Nie, L.; Zhao, Y.-L.; Akbari, M.; Shen, J.; and Chua, T.-S. 2015. Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE TKDE* 27(2):396–409.
- 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.
- Tay, Y.; Anh Tuan, L.; and Hui, S. C. 2018. Latent relational metric learning via memory-based attention for collaborative ranking. In *WWW*, 729–739.
- 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.; Belongie, S.; and Karaletsos, T. 2017. Conditional similarity networks. In *CVPR*, 1781–1789. IEEE.
- 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.
- Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*, 1386–1393.
- Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, 2659–2665.
- Xie, R.; Liu, Z.; and Sun, M. 2016. Representation learning of knowledge graphs with hierarchical types. In *IJCAI*, 2965–2971. AAAI Press.
- Yang, X.; Wang, M.; and Tao, D. 2018. Person reidentification with metric learning using privileged information. *IEEE 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.
- Yu, W.; Zhang, H.; He, X.; Chen, X.; Xiong, L.; and Qin, Z. 2018. Aesthetic-based clothing recommendation. In *WWW*, 649–658.
- Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *CVPR*, 3107–3115. IEEE.