Fashion Recommendation with Multi-relational Representation Learning

Yang Li, Yadan Luo, and Zi Huang(✉)

School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, QLD 4072, Australia
yang.li@uq.edu.au, lyadanluol@gmail.com, huang@itee.uq.edu.au

Abstract. Driven by increasing demands of assisting users to dress and match clothing properly, fashion recommendation has attracted wide attention. Its core idea is to model the compatibility among fashion items by jointly projecting embedding into a unified space. However, modeling the item compatibility in such a category-agnostic manner could barely preserve intra-class variance, thus resulting in sub-optimal performance. In this paper, we propose a novel category-aware metric learning framework, which not only learns the cross-category compatibility notions but also preserves the intra-category diversity among items. Specifically, we define a category complementary relation representing a pair of category labels, e.g., tops-bottoms. Given a pair of item embeddings, we first project them to their corresponding relation space, then model the mutual relation of a pair of categories as a relation transition vector to capture compatibility amongst fashion items. We further derive a negative sampling strategy with non-trivial instances to enable the generation of expressive and discriminative item representations. Comprehensive experimental results conducted on two public datasets demonstrate the superiority and feasibility of our proposed approach.

Keywords: Fashion compatibility · Fashion recommendation · Representation learning

1 Introduction

With the proliferation of online fashion websites, such as Polyvore[1] and Farfetch[2], there are increasing demands on intelligent applications in the fashion domain for a better user shopping experience. This drives researchers to develop various machine learning techniques to meet such demands. Existing work is mainly conducted for three types of fashion applications: (1) clothing retrieval [1,1,8]: retrieving similar clothing items from the data collection based on the query clothing item; (2) fashion attribute detection [3,11,12]: identifying clothing attributes such as color, pattern and texture from the given clothing image;

[1] www.polyvore.com.
[2] www.farfetch.com.

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 3–15, 2020.
https://doi.org/10.1007/978-3-030-47426-3_1

(3) Complementary Clothing Recommendation [5,10,16,21,22]: recommending complementary clothes that match the query clothing item to the user. In this paper, we focus on the third application, which is more challenging and sophisticated due to the fashion data complexity and heterogeneity. It requires the model to infer compatibility among fashion items according to various complementary characteristics, which goes beyond visual similarity measurement.

The key point to tackle the above challenges is to derive an appropriate compatibility measurement for pairs of fashion items, which can effectively capture various fashion attributes (e.g., colors and patterns) from item images for comparison. The major stream of existing approaches for fashion compatibility modeling adopts metric learning techniques to extract effective fashion item representations. A typical fashion compatibility modeling strategy is to learn a latent style space, where matching item pairs stay closer than incompatible pairs. The compatibility of two given fashion items is computed by the pairwise Euclidean distance or inner product between fashion item embeddings. Nevertheless, the previous work has two main limitations that lead to sub-optimal performance. Firstly, some approaches consider fashion compatibility modeling as a single-relational task. However, this neglects the fact that people usually focus on different aspects of clothes from different categories. For example, people are more likely to focus on color and material for blouses and pants, while they may pay attention to shape and style for jeans and shoes. Moreover, using a single unified space is likely to result in incorrect similarity transitivity in fashion compatibility. For instance, if item A matches both B and C, while B and C may not be compatible, the embeddings of A, B and C will be forced to be close to each other in a single unified space, which degrades prediction performance because the compatibility essentially does not hold transitivity property. Therefore, such a category-independent approach will result in inaccurate item representations. Secondly, most existing approaches merely randomly sample negative instances from the training set. However, most of the randomly sampled triplets are trivial ones, which may fail to support the model to learn discriminative item representations.

In order to address the above mentioned limitations, we propose a novel **C**ategory-**A**ware **F**ashion **M**etric **E**mbedding learning network (CA-FME), which models both instances and category-aware relation representations through a translation operation. Specifically, we formulate the fashion compatibility measurement as a multi-relational data modeling task. We treat fashion items as entities and define pairs of compatible categories as complementary relations, e.g., blouses-skirts. The overall flowchart of CA-FME is presented in Fig. 1: Item visual features are first extracted through a pre-trained CNN. Then, each pair of item embeddings is projected to their corresponding category-specific relation subspace. Finally, we model the compatibility based on a transition-based score function. Our main contributions can be summarized as below:

– We present a novel category-aware embedding learning framework for fashion compatibility modeling, which not only captures cross-categorical relationships but also preserves the diversity of intra-category fashion item representations.

– We devise a negative sampling strategy with non-trivial samples for discriminative item representations.
– Extensive experiments have been conducted on two real world datasets, Polyvore and FashionVC, to demonstrate the superior performance of our model over other state-of-the-art methods.

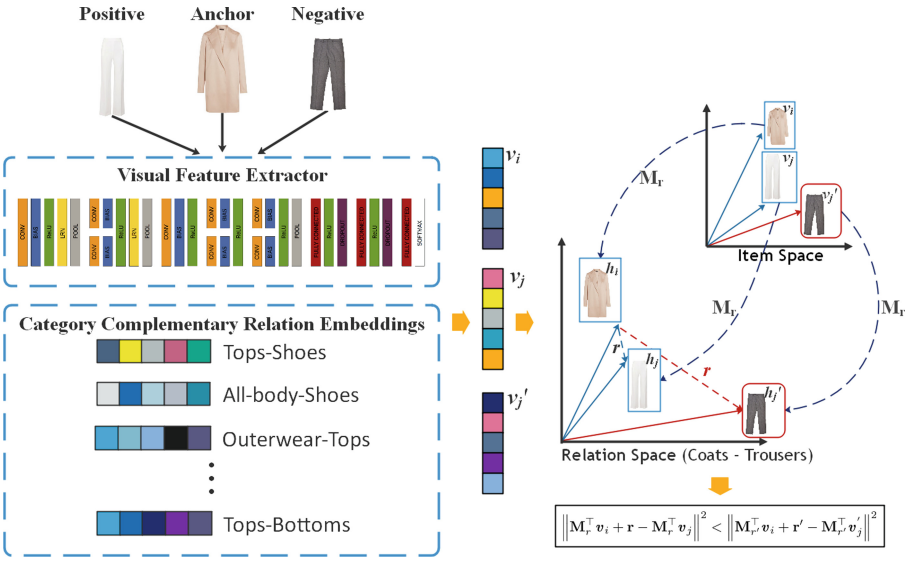

Fig. 1. The overview of proposed CA-FME model architecture for fashion compatibility modeling. The fashion clothing dataset consists of multiple categories, e.g., Hoodies, Skirts, Coats and Trousers. CA-FME mainly consists of three parts: (1) A pre-trained CNN for visual feature extraction; (2) A category complementary relation embedding space for modeling category-aware compatibility; (3) Multiple relation-specific projection spaces for preserving the intra-class diversity. The whole framework is finally optimized via a margin-based ranking objective function in end-to-end manner.

2 Related Work

2.1 Fashion Compatibility Modeling

The mainstream of work aims to map fashion items into a latent space where compatible item pairs are close to each other, while incompatible pairs lay in the opposite position. McAuley et al. [13] propose to use Low-rank Mahalanobis Transformation to learn a latent style space for minimizing the distance between matched items and maximizing that of mismatched ones. Following this work, Veit et al. [19] employ the Siamese CNNs to learn a metric for compatibility measurement in an end-to-end manner. Some researchers argue that the complex compatibility cannot be captured by directly learning a single latent space.

He et. al [6] propose to learn a mixture of multiple metrics with weight confidences to model the relationships between heterogeneous items. Veit et al. [18] propose Conditional Similarity Network, which learns disentangled item features whose dimensions can be used for separate similarity measurements. Following this work, Vasileva et al. [17] claim that respecting type information has important consequences. Thus, they first form type-type spaces from each pair of types and train these spaces with triplet loss.

2.2 Knowledge Graph Embedding Learning

The techniques of representation learning on the knowledge graph have attracted large attention in recent years. Different from the approaches implemented by tensor factorization, e.g., [14], translation-based models [2,7,20], which is partially inspired by the idea of word2vec, have achieved state-of-the-art performance in the field of the knowledge graph. Similar to the knowledge graph, heterogeneous fashion recommendation can also be considered as a multi-relational problem, where complementary categories form various relations. Enlightened by these findings, we apply a similar idea from the knowledge graph to the fashion domain for compatibility modeling.

3 Problem Formulation

The fashion complementary recommendation task we are tackling is formulated as follows. Suppose we have a collection of fashion item images denoted as $\mathcal{O} = \{o_1, o_2, o_3, ..., o_n\}$, where n is the number of items, and a set of category labels denoted as $\mathcal{C} = \{c_1, c_2, c_3, ..., c_m\}$, where m is the number of categories. Each fashion item $o_i \in \mathcal{O}$ has a corresponding k-dim visual feature vector $\boldsymbol{v}_i = g(o_i; \Theta_v), \boldsymbol{v}_i \in \mathbb{R}^k$ and a category label $c_i \in \mathcal{C}$. Here, $g(o; \Theta_v)$ represents a pre-trained CNN with trainable parameters Θ_v, which extracts visual features from a fashion item image $o \in \mathcal{O}$. We denote a set of category complementary relations as $\mathcal{R} = \{r^{c_i c_j}\}$, where $c_i, c_j \in \mathcal{C}$ represent a pair of complementary categories, such as *tops-bottoms*. We now use a triplet $(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j}), s.t., \forall i, j, r^{c_i c_j} \in \mathcal{R}$ to represent embeddings of a pair of fashion items o_i and o_j and their corresponding category complementary relation $r^{c_i c_j}$. Each relation $r^{c_i c_j} \in \mathcal{R}$ corresponds to an embedding vector $\boldsymbol{r}^{c_i c_j} \in \mathbb{R}^d$ from the relation embedding space. Our target is to derive a fashion compatibility scoring function $f(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j})$, which captures visual characteristics from the item embeddings for compatibility measurement.

4 Proposed Approach

In this section, we first present our CA-FME model for fashion compatibility modeling. Then, we introduce a novel negative sampling strategy for more effective training. Finally, we describe the optimization algorithm to train our model.

The overview of our proposed framework is shown in Fig. 1. We aim to build a model, which can (1) effectively model the notion of compatibility; (2) be easily generalized to unseen fashion item compatibility measurement; (3) focus on different aspects of item embeddings regarding different category complementary relations for the compatibility measurement. In particular, the framework consists of a pre-trained CNN for visual feature extraction and multiple category complementary relation subspaces for category-aware compatibility modeling.

4.1 Compatibility Modeling

To solve the above mentioned limitations, we assign each category complementary relation $r \in \mathcal{R}$ with a single d-dim transition vector $\boldsymbol{r} \in \mathbb{R}^d$. Intuitively, these relation vectors act as different fashion compatibility decision-makers who focus on different pairs of categories, which enables the model to concentrate on different aspects of fashion items from different categories. In particular, given a pair of fashion items o_i and o_j with their visual features $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, and their corresponding category complementary relation $r^{c_i c_j}$. If o_i is compatible with o_j, the compatibility relationship can be interpreted as:

$$\boldsymbol{v}_i + \boldsymbol{r}^{c_i c_j} \approx \boldsymbol{v}_j \tag{1}$$

which means o_j's embedding $\boldsymbol{v}_j$ should be the nearest neighbor to the resulting vector of $\boldsymbol{v}_i$ plus the relation vector $\boldsymbol{r}^{c_i c_j}$ in a specific latent space based on a certain distance metric, e.g., L1 or L2 distance.

However, there exists one issue in the above equation: in reality, items from a specific pair of categories share diverse fashion attributes such as material, style and pattern. Therefore, it is insufficient to preserve intra-category diversity by building only a single embedding vector for each category complementary relation. To address this issue, we propose to build multiple relation-specific subspaces, i.e., $\mathbf{M}_r \in \mathbb{R}^{k \times d}, r \in \mathcal{R}$, where k is the number of visual feature vector dimensions. Using such category-aware projection operations is twofold. Firstly, the relation-specific subspaces provide abundant trainable parameters to preserve intra-category diversity. Secondly, it also provides capability for handling unseen items through a projection operation. Thus, we define the projected item vectors of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ as,

$$\boldsymbol{h}_i = \mathbf{M}_{r^{c_i c_j}}^{\top} \boldsymbol{v}_i, \quad \boldsymbol{h}_j = \mathbf{M}_{r^{c_i c_j}}^{\top} \boldsymbol{v}_j, \quad \boldsymbol{h}_i, \boldsymbol{h}_j \in \mathbb{R}^d \tag{2}$$

With the above defined compatibility relationship modeling rule and relation-specific projection, we now could perform compatibility score calculation within the corresponding relation space. Given a pair of fashion items denoted as o_i and o_j, and their corresponding category complementary relation $r^{c_i c_j}$, the compatibility score s_{ij} is calculated as,

$$s_{ij} = - \left\| \boldsymbol{h}_i + \boldsymbol{r}^{c_i c_j} - \boldsymbol{h}_j \right\|_2 \tag{3}$$

where L2 distance is used.

Algorithm 1: Negative Sampling

Input : $(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j})$: a positive triplet,
$\hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$: negative candidate set,
$\bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$: selected negative triplet set,
N: size of negative candidate set,
M: size of selected negative triplet set, $M < N$

Output: The set of M negative training triplets $\bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$

1 Construct negative candidate set $\hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})} = \{\boldsymbol{v}'_1, \boldsymbol{v}'_2, ..., \boldsymbol{v}'_N\}$, where $(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j}) \in \mathcal{N}$ and $c'_j = c_j$ by uniformly sampling.

2 Compute the score $f(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j})$ for all $\boldsymbol{v}'_j \in \hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$ via Equation (3)

3 Construct the selected negative triplet set $\bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$ by multinomial sampling M items from $\hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$ with probability in Equation (6)

4 **Return:** $\bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$

4.2 Negative Sampling

Negative sampling has been proven to be an effective and helpful training strategy to learn discriminative item representations in various fields. We aim to derive a simple but effective negative sampling strategy to assist our model to identify more subtle style patterns from hard negative instances. Since a category complementary relation corresponds to two different categories, we want both sides of each training triplet can benefit from negative sampling. Therefore, we define the strategy should meet the following requirements:

1. The strategy should consider both sides of training triplets.
2. The strategy should identify hard negative instances effectively and efficiently.
3. The strategy should avoid false negative samples effectively.

Now we introduce the details regarding how our designed negative sampling strategy can meet the above-defined requirements. We also present the details of our strategy in Algorithm 1.

Requirement 1: We propose to sample negative instances from both sides of a given positive triplet $(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j})$. In particular, we first fix $\boldsymbol{v}_i$ and category complementary relation $r^{c_i c_j}$, then replace $\boldsymbol{v}_j$ by randomly sampling an item embedding vector $\boldsymbol{v}'_j$ from category c_j. Similarly, we perform the same negative sampling for the other side item $\boldsymbol{v}_j$.

Requirement 2: Given a positive triplet $(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j})$, we first uniformly sample N negative candidates denoted as $\hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$ from category c_j's item set. Then, for each training triplet, we calculate scores for all negative triplets. This two steps correspond to the step 1–2 in Algorithm 1. Intuitively, the negative triplets with high compatibility scores can be regarded as hard negative samples.

Requirement 3: Despite the higher scores the harder negative samples are, these samples are likely to be false negative, which instead has destructive impact on the model performance. In order to avoid this issue, we propose to select

Algorithm 2: Training CA-FME

Data : Training set of positive triplets $\mathcal{P}$, negative triplets $\mathcal{N}$
Input: $g(o;\theta_v)$: pre-trained CNN with parameters θ_v for visual feature extraction,
$\mathcal{R}$: category relation set,
$\bar{\mathcal{H}}$: negative triplets,
$\tilde{\mathcal{H}}$: 4-tuple training set,
B: batch size

initialize $\boldsymbol{r}$ by Xavier initialization for each $r \in R$,
$\mathbf{M}_r$ by Xavier initialization for each $r \in R$
repeat
  Sample a training batch S_{batch} from $\mathcal{P}$ with batch size B
  $T_{batch} \leftarrow \varnothing$
  for $(o_i, o_j, r^{c_i c_j}) \in S_{batch}$ **do**
    $\boldsymbol{v}_i = g(o_i;\theta_v)$
    $\boldsymbol{v}_j = g(o_j;\theta_v)$
    `// Get negative triplets with` v_i `and` $r^{c_i c_j}$ `fixed`
    Construct negative triplets $\bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})} = \{(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j})\}$ via Algo. 1
    Form the 4-tuple training set
    $\tilde{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})} = \{(\boldsymbol{v}_i, \boldsymbol{v}'_i, \boldsymbol{v}_j, r^{c_i c_j})\}, \boldsymbol{v}'_i \in \bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$
    `// Get negative triplets with` v_j `and` $r^{c_i c_j}$ `fixed`
    Construct negative triplets $\bar{\mathcal{H}}_{(\boldsymbol{v}_j, r^{c_i c_j})} = \{(\boldsymbol{v}'_i, \boldsymbol{v}_j, r^{c_i c_j})\}$ via Algo. 1
    Form the 4-tuple training set
    $\tilde{\mathcal{H}}_{(\boldsymbol{v}_j, r^{c_i c_j})} = \{(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{v}'_j, r^{c_i c_j})\}, \boldsymbol{v}'_j \in \bar{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$
    $T_{batch} \leftarrow T_{batch} \cup \tilde{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})} \cup \tilde{\mathcal{H}}_{(\boldsymbol{v}_j, r^{c_i c_j})}$
  endfor
  Update the whole network via Hinge loss function:
  $\sum_{T_{batch}} \nabla[\gamma + f(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j}) - f(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j})]$
until *Convergence*;

M negative items from the above sampled N negative candidates with different probability by multinomial sampling, which corresponds to step 3 in Algorithm 1. In particular, we grant larger probability for harder negative samples according to their scores. Here, let $\mathcal{S} = \{s_1, s_2, ..., s_N\}$ be the set of calculated scores of N negative candidates. We first define the following normalization function $norm(s_{ij})$ to project all the scores into the range of $[0, 1]$,

$$norm(s_{ij}) = \frac{s_{ij} - s_{min}}{s_{max} - s_{min}}, s_{ij} \in \mathcal{S}, s_{min} = min(\mathcal{S}), \tag{4}$$

$$s_{max} = max(\mathcal{S}) \tag{5}$$

Finally, we could define the probability of sampling a negative item $\bar{\boldsymbol{v}}$ by:

$$p(\bar{\boldsymbol{v}}_j | (\boldsymbol{v}_i, r^{c_i c_j})) = \frac{\exp(1 - norm(f(\boldsymbol{v}_i, \bar{\boldsymbol{v}}_j, r^{c_i c_j})))}{\sum_{(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j}) \in \hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}} \exp\left(1 - norm(f\left(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j}\right))\right)} \tag{6}$$

Margin-Based Optimization. With the above defined score function and negative sampling strategy, we present the whole training steps in Algorithm 16. Let $\hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}$ and $\hat{\mathcal{H}}_{(\boldsymbol{v}_j, r^{c_i c_j})}$ denote the 4-tuple training triplets constructed using the above defined negative sampling strategy. We could define the following margin-based loss function as our objective function for training:

$$\begin{aligned}\mathcal{L} = & \sum_{\left(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{v}'_j, r^{c_i c_j}\right) \in \hat{\mathcal{H}}_{(\boldsymbol{v}_i, r^{c_i c_j})}} \left[f\left(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j}\right) - f\left(\boldsymbol{v}_i, \boldsymbol{v}'_j, r^{c_i c_j}\right) + \gamma\right]_+ + \\ & \sum_{\left(\boldsymbol{v}_i, \boldsymbol{v}'_i, \boldsymbol{v}_j, r^{c_i c_j}\right) \in \hat{\mathcal{H}}_{(\boldsymbol{v}_j, r^{c_i c_j})}} \left[f\left(\boldsymbol{v}_i, \boldsymbol{v}_j, r^{c_i c_j}\right) - f\left(\boldsymbol{v}'_i, \boldsymbol{v}_j, r^{c_i c_j}\right) + \gamma\right]_+\end{aligned} \tag{7}$$

where γ is the margin value and $[x]_+ \triangleq \max(0, x)$.

We adopt the stochastic gradient decent algorithm (SGD) for the model optimization. In each step, we sample a mini-batch of training triplets and update the parameters of the whole network.

5 Experiments

In this section, we first describe the experimental settings and then give comprehensive analysis based on the experimental results.

5.1 Dataset

We conduct our experiments on two public datasets, FashionVC and Polyvore-Maryland, provided by Song et al. [16] and Han et al. [5] respectively.

FashionVC [16]**.** This dataset consists of 14,871 top item images and 13,663 bottom item images, where each item has a corresponding image, a title and a category label. In this paper, we only consider the visual modality. Therefore, we use images for visual information extraction and category labels to determine which category complementary relation the item pairs belong to. We randomly split the data according to 80%;10%;10% for training, validation and test sets, respectively.

PolyvoreMaryland [5]**.** This dataset contains 21,799 outfits crawled from the online social community website Polyvore. We use the splits provided by Han et al. [5], which has 17,316, 3,076 and 1,407 outfits in training, testing and validation sets respectively. In this paper, we mainly study item-to-item compatibility, therefore, we keep four main groups of fashion items: tops, bottoms, bags and shoes from the outfit data. Each fashion item contains an image, a title and a category label. Note that each group of fashion items have several detailed category labels, e.g., there are hand bags and shoulder bags in the "bags" group.

5.2 Baseline Methods

We compare our model CA-FME with several state-of-the-art models for heterogeneous recommendation. For the fair comparison, we set the pre-trained Alexnet [9] as the visual feature extractor of all methods.

- **SiameseNet** [19]: The approach models compatibility by minimizing the Euclidean distance between compatible pairs and maximizing the distance between incompatible ones in a unified latent space through contrastive loss.
- **Monomer** [6]: The approach models fashion compatibility with a mixture of distances computed from multiple latent spaces.
- **BPR-DAE** [16]: The approach models compatibility through inner-product result of top's and bottom's embeddings and uses Bayesian Personalized Ranking (BPR) [15] as their optimization objective.
- **TripletNet** [4]: The approach models fashion compatibility in a unified latent space through triplet loss.
- **TransNFCM** [22]: The state-of-the-art method that learns item-item compatibility by modeling categorical relations among different fashion items.
- **TA-CSN** [17]: The state-of-the-art method that builds type-aware subspaces for fashion compatibility modeling.

5.3 Parameter Settings

In our experiment, all the hyper-parameters of our approach are tuned to perform the best on the validation set. For the fair comparison, we apply the Alexnet [9] as the visual feature extractor for all methods. In our model, we set margin γ as 1, learning rate $\alpha = 10^{-4}$ with momentum 0.9, batch size $B = 512$. Visual embedding dimension $k = 128$, with dropout rate 0.5 and relation embedding dimension is set to be 128.

5.4 Compatibility Prediction

Task Description. The compatibility prediction task aims to predict whether a given pair of items are compatible or not. In particular, we replace one item of each testing positive triplet with 100 randomly sampled negative items. Thus, for each testing instance, it requires to give ranking on 101 items based on the query image. We employ two widely-used evaluation metrics, Hit@k and Area Under the ROC curve (AUC) to evaluate the performance of our model and baseline methods based on the predicted compatibility scores. Hit@k is defined as follows, which indicates the proportion of the correct predicted item ranked in top k.

$$\text{Hits@}k = \frac{\#\text{hit@}k}{\|D_{\text{test}}\|} \tag{8}$$

where D_{test} denotes the collection of testing instances. The formula for AUC is defined as below,

$$AUC = \frac{\sum pred_{positive} > pred_{negative}}{|positiveInstances| \times |negativeInstances|} \tag{9}$$

where $\sum pred_{positive} > pred_{negative}$ indicates the number of cases that the predicted score of positive instance is larger than negative one, by comparing the predicted score of each positive instance with each negative instances in the testing set.

Table 1. Performance comparison between our proposed CA-FME and other baseline methods. CA-FME(Neg.) indicates the application of negative sampling training strategy.

FashionVC						PolyvoreMaryland				
Methods	AUC	Hit@5	Hit@10	Hit@20	Hit@40	AUC	Hit@5	Hit@10	Hit@20	Hit@40
SiameseNet	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
BPR-DAE	70.9	16.7	27.3	46.7	70.4	69.5	17.3	28.2	43.9	67.5
Triplet Net	70.6	16.3	28.0	45.7	69.6	70.1	18.1	28.7	44.9	68.3
TA-CSN	71.6	16.7	28.4	46.7	70.8	70.2	17.3	28.4	45.1	68.4
TransNFCM	73.6	19.0	32.3	51.6	74.0	73.6	19.3	33.1	50.9	73.4
CA-FME	88.6	**26.6**	48.5	81.9	**99.9**	95.0	**59.8**	84.4	**97.7**	**99.7**
CA-FME (Neg.)	**88.9**	26.4	**49.9**	**83.2**	**99.9**	**96.2**	59.6	**88.4**	96.7	**99.7**

5.5 Performance Comparison

We evaluate our model with and without negative sampling strategy, i.e., CA-FME(Neg) and CA-FME. Table 1 shows the performance comparison on two datasets based on AUC and Hit@K evaluation metrics. From the table we have the following observations:

- Our model achieves the best performance on both datasets by significant margins compared with all the other state-of-the-art methods, which proves the effectiveness and superior performance of our method.
- The category-unaware models including SiameseNet and TripletNet, which merely learn fashion compatibility notions in a single latent space, perform worse than category-respected models including TA-CSN and TransNFCM. This proves that considering category label information is of great importance in fashion compatibility modeling, which can be helpful to avoid incorrect compatibility similarity transitivity. It also proves that items from different categories may have very different visual characteristics for compatibility.
- Compared with category-aware methods, TA-CSN and TransNFCM, our model obtains around 15% and 30% improvements on AUC and Hit@20 respectively. Although they build category-aware mask vectors to capture different fashion characteristics among different categories, it is still not sufficient to preserve the intra-category diversity among items. With the help of our relation-specific projection spaces, our model can capture much more specific information of compatibility from different categories. The improvements on PolyvoreMaryland dataset are even much better in terms of AUC and Hit@5. This is mainly because of the different number of relations in two

datasets. We define 146 category relations in the Polyvore dataset, while there are only 30 relations in the FashionVC dataset. It proves that more relational spaces can significantly contribute to the improvement of performance.
– The results of CA-FME(Neg.) show that our negative sampling strategy is helpful to improve our model's performance, which proves the effectiveness of our proposed training strategy.

5.6 Case Study

In this section, we conduct a case study, aiming to address a real-world fashion recommendation task: selecting the fashion item that matches the query one. As illustrated in Fig. 2, we conduct two query instances on the FashionVC dataset, where the items with a green box are ground-truths. In the first case, we give the model a woman blouse, the model successfully selects the ground-truth at first rank. It can be observed that the model identifies the color of the first ranked jeans matches the query blouse. Our model also successfully identifies that the 7^{th} jeans are for men and thus gives it the lowest score. In the second case, the model gives a relatively high score to the ground-truth item. However, the main reason that our model gives a higher score to the first item probably due to the color attribute. For the latter items ranked at 5–7, we think our model successfully identifies that their shapes do not match the query skirt.

Fig. 2. Case study of fashion recommendation task by retrieving the most matching fashion items from a set of candidates based on the query fashion item. (Color figure online)

6 Conclusion

In this work, we introduced a novel category-aware neural model CA-FME to model the fashion compatibility notions. It not only captures cross-category compatibility by constructing category relation embeddings but also preserves intra-category diversity among items through build relation-specific projection spaces. To optimize our model, we further introduce a weighted negative sampling strategy to identify high-quality negative instances, which consequently

assists our model to infer discriminative representations. In addition, although in our paper, we mainly study the compatibility of tops and bottoms, it can easily generalized to arbitrary types of clothing items. Extensive experiments were conducted on two public fashion datasets, which shows that our CA-FME model can significantly outperform all the state-of-the-art methods on fashion recommendation.

Acknowledgments. We would like to thank all reviewers for their comments. This work was partially supported by Australian Research Council Discovery Project (ARC DP190102353).

References

1. Ak, K.E., Kassim, A.A., Lim, J.H., Tham, J.Y.: Learning attribute representations with localization for flexible fashion search. In: CVPR (2018)
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS, pp. 2787–2795 (2013)
3. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_44
4. Chen, L., He, Y.: Dress fashionably: learn fashion collocation with deep mixed-category metric learning. In: AAAI, pp. 2103–2110 (2018)
5. Han, X., Wu, Z., Jiang, Y., Davis, L.S.: Learning fashion compatibility with bidirectional LSTMS. In: ACM MM, pp. 1078–1086 (2017)
6. He, R., Packer, C., McAuley, J.J.: Learning compatibility across categories for heterogeneous item recommendation. In: ICDM, pp. 937–942 (2016)
7. Ji, G., Liu, K., He, S., Zhao, J.: Knowledge graph completion with adaptive sparse transfer matrix. In: AAAI (2016)
8. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: matching street clothing photos in online shops. In: ICCV (2015)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
10. Li, Y., Luo, Y., Huang, Z.: Graph-based relation-aware representation learning for clothing matching. In: Borovica-Gajic, R., Qi, J., Wang, W. (eds.) ADC 2020. LNCS, vol. 12008, pp. 189–197. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39469-1_15
11. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: CVPR. IEEE (2016)
12. Luo, Y., Wang, Z., Huang, Z., Yang, Y., Zhao, C.: Coarse-to-fine annotation enrichment for semantic segmentation learning. In: CIKM (2018)
13. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR, pp. 43–52 (2015)
14. Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In: ICML, pp. 809–816 (2011)
15. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: bayesian personalized ranking from implicit feedback. In: IJAI, pp. 452–461 (2009)

16. Song, X., Feng, F., Liu, J., Li, Z., Nie, L., Ma, J.: Neurostylist: neural compatibility modeling for clothing matching. In: ACM MM, pp. 753–761 (2017)
17. Vasileva, M.I., Plummer, B.A., Dusad, K., Rajpal, S., Kumar, R., Forsyth, D.: Learning type-aware embeddings for fashion compatibility. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part XVI. LNCS, vol. 11220, pp. 405–421. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_24
18. Veit, A., Belongie, S.J., Karaletsos, T.: Conditional similarity networks. In: CVPR, pp. 1781–1789 (2017)
19. Veit, A., Kovacs, B., Bell, S., McAuley, J.J., Bala, K., Belongie, S.J.: Learning visual clothing style with heterogeneous dyadic co-occurrences. In: ICCV, pp. 4642–4650 (2015)
20. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI (2014)
21. Yang, X., et al.: Interpretable fashion matching with rich attributes. In: SIGIR (2019)
22. Yang, X., Ma, Y., Liao, L., Wang, M., Chua, T.: TransNFCM: translation-based neural fashion compatibility modeling. In: AAAI, pp. 403–410 (2019)