Content-aware generative model for multi-item outfit recommendation

Valery Volokha[1] and Klavdiya Bochenina[1]

[1] ITMO University, St. Petersburg Kronverksky Pr. 49 bldg. A 197101, Russia

Abstract. Recently, deep learning-based recommender systems have received increasing attention of researchers and demonstrate excellent results at solving various tasks in various areas. One of the last growing trends is learning the compatibility of items in a set and predicting the next item or several ones by input ones. Fashion compatibility modeling is one of the areas in which this task is being actively researched. Classical solutions are training on existing sets and are learning to recommend items that have been combined with each other before. This severely limits the number of possible combinations. GAN models proved to be the most effective for decreasing the impact of this problem and generating unseen combinations of items, but they also have several limitations. They use a fixed number of input and output items. However, real outfits contain a variable number of items. Also, they use unimodal or multimodal data to generate only visual features. However, this approach is not guaranteed to save content attributes of items during generation. We propose a multimodal transformer-based GAN with cross-modal attention to simultaneously explore visual features and textual attributes. We also propose to represent a set of items as a sequence of items to allow the model to decide how many items should be in the set. Experimenting on FOTOS dataset at the fill-in-the-blank task is showed that our method outperforms such strong baselines as Bi-LSTM-VSE, MGCM, HFGN, and others. Our model has reached 0.878 accuracy versus 0.724 of Bi-LSTM-VSE, 0.822 of MGCM, 0.826 of HFGN.

Keywords: Outfit Recommendations, Set Recommendations, Multimodal Recommendations, Generative Adversarial Networks (GAN), Transformers, Recommender Systems.

1 Introduction

In recent years e-commerce has spread widely, especially under the influence of COVID-19 and related restrictions. A lot of people turned to online shopping over personal visits to shops, which led to an unbound variety of items to compare and combine during the shopping process. The fashion industry and online fashion marketplaces are one of the areas largely affected by this. Fashion compatibility modeling is of increasing interest for researchers and becomes a popular but challenging and contentious topic. There are a lot of downstream applications such as the outfit recommendation [1–11], the personalized fashion design [12–16], personal wardrobe creation [17,18], fashion-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

oriented dialogue systems [19,20], try-on [19,20], and others. In this paper, we combined the task of personalized fashion design generation with an outfit recommendations task. We used a generative model to generate new personalized item representations but used them to find real equivalents to build outfit recommendations.

The rapid development of technologies and the increased computational power in the last decade allowed recommender systems to integrate into various areas of our life, including the fashion industry and e-commerce. Modern shopping apps assist and influence customer decisions. Therefore, this is becoming increasingly important to develop personalized and efficient recommender systems for choosing a set of clothes. The main aim of these systems is to automatically assess the compatibility of items and predict missing items of outfits. This area is actively researched, and there is impressive progress, but there are still has some unsolved problems, which limits the efficiency and the flexible usage of these systems.

Classical recommender systems learn to recommend items that have been already combined with input ones before, but it severely limits the compatibility of items and the variety of outfits. Several approaches tried to decrease the impact of this problem. Some of them applied noise to vector representations of input items. Others used variational autoencoders as a base of a model. They showed the effectiveness of recommendations but did not inspect the ability of the model to recommend items that had not been previously combined with the input ones. They only reduced the discontinuity of the latent space of the model, but it is not a complete solution, and the model is still fitted to recommend existing outfits but not to generate the most compatible items.

Generative adversarial network (GAN) based models are used to overcome this problem [12–16] but they also have some limitations. In particular, they use a fixed number of input and output items (primarily, one or two input items and a single output item) [12–16]. The main reason is that they frequently use noise as a placeholder for blanked items. A large amount of input noise makes the model unpredictable and reduces the influence of the input items to newly generated ones. Moreover, to the best of our knowledge, the presented GAN-based solutions aim to synthesize images of new items [12–16], but in the case of e-commerce, online shops, and recommending existing items that users can buy, there is no need to generate images directly.

Consequently, in this paper, we have focused on the compatibility modeling sets with a variable number of items by data of multiple modalities. Despite the fact that the data from several modalities are used in many approaches, to the best of our knowledge, most of these explore modalities separately in the field of recommending sets of items, including in the field of fashion. On the contrary, in our scheme, we have focused on capturing compatibility features between modalities simultaneously.

In this paper, we propose the following contributions to solve described problems:

- Different from the existing GAN-based methods which have a fixed number of input and output items, we propose to use a transformer-based GAN and represent a set of items as a sequence of items with start and end tags, similar to a sentence of words. This allows the model not only to generate a complete set of items but also to decide how many items should be in the set.
- To simultaneously explore multimodal features, we propose to use a cross-modal attention module in our transformer. Transformer architecture with self-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

attention and cross-modal attention allows GAN to jointly generate vectors of visual and textual features depending on multimodal input ones.

The rest of the paper is organized as follows. First, we described and discussed several related recommender-based and GAN-based studies. Second, we formulated a problem and described our proposed model and its parts. Then, we described conducted experiments, used dataset, compared with our model baselines, and comparison results. Finally, we summarized the contributions of this paper and obtained results.

2 Related works

2.1 Recommender-based solutions

Plummer B.A. et. al. (2018) proposed the method that embeds compatible items and outfits close to each other for searching similar items to the input ones and replacing items if necessary [21]. Tangsend P. et. al. (2018) also embedded items and used a binary classifier to predict the compatibility of items within a set and to rank sets by compatibility [4]. Lu Z. et. al. (2019) used CNN to extract visual features from images of all types of items, and type-specific embeddings (each item type is associated with its own embedding) to project feature vectors depending on the type of item [7]. The authors proposed a fashion hashing network (FHN) which uses HashNet and BPR to assess the compatibility of computed feature vectors of a set. They personalize recommendations by adding a vector of user features. The problem with such approaches is that they only use visual features of items and ignore content attributes and descriptions. However, content information is very important and is near-always available in real conditions. For example, in the outfit recommendation task, black jeans and black leggings are close to each other in embedding space, but they are very different, and such content attribute as a material can help to separate them.

To overcome this problem, Xintong H. et. al. (2017) proposed multimodal Bi-LSTM to sequentially predict the next item conditioned on previous ones to learn their compatibility relationships [22]. They also proposed a method to explore visual and textual together by projecting visual features to the space of textual attributes.

Cui Z. et. al. (2019) proposed multimodal graph-based neural network that optimizes Fashion Graph and uses the attention layer to compute the compatibility score [8]. The authors used deep convolutional neural network to extract visual features from images and one-hot-encoding to represent titles as boolean vector. They described a strategy to train the multimodal node-wise graph neural network (NGNN) and showed that their approach outperforms such baselines as unimodal GGNN, Bi-LSTM, and others.

Li X. et. al. (2020) also proposed multi-model graph-based neural network [6] and unified two tasks: fashion compatibility modeling and personalized outfit recommendation. They proposed Hierarchical Fashion Graph Network (HFGN) to model relationships among users, items and outfits simultaneously. They also proposed an R-view attention map, which can capture the potential compatibility knowledge better. The authors demonstrated that their model outperforms such state-of-the-art methods as NGNN and FHN.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

Cardoso A. et. al. (2018) proposed multimodal embedding that takes item images, type, description and some content characteristics and projects them to interpreted categorical embedding [9]. The authors also introduced a hybrid architecture to compare and rank items that combines content-based and collaborative inputs as well as an embedding to project them. Elaine M.B. et. al. (2019) proposed a method to assess the compatibility of items that use this embedding [23]. They showed how to recommend a highly scored set of items to a user by several input items. Their approach uses embeddings and applies dot product and softmax operations to calculate the compatibility score.

Sagar D. et. al. (2020) proposed a multi-model method to personalized outfit recommendations with attribute-wise interpretability [10]. Their method is based on BPR and ranks triplets of items. The obtained vectors of features are integrated with the embedded user preferences vector and used by BPR to compute the compatibility score. The authors showed that their method outperforms such baselines as Bi-LSTM, VTBPR, GP-BPR, and others.

Yuan F. et. al. (2018) proposed simple convolutional generative network for next item recommendation based on dilated CNN architecture [11]. The authors tried to implement the idea of learning short- and long-range dependencies between items using CNN. They stacked holed convolutional layers and used residual block structure. Results showed that the proposed generative model attains state-of-the-art accuracy.

2.2 GAN-based solutions

The main problem with the approaches proposed above is that they are trained to recommend items that were encountered in existed sets along with the items received as input. This severely limits the possible variety of combinations of items. Some authors tried to solve this problem by applying noise to the input items or using a variational autoencoder as the base of their model, but it does not solve the problem completely, but makes the latent space of the model less sparse. Compatible items that are not presented in the existing sets will still not be recommended together. To overcome this problem, it is proposed to use generative adversarial neural networks (GAN) that explore the items and the compatibility of items and learn to generate new items from noise. In order to offer real items to the user, the generated items are compared with the real ones from the target dataset.

Kang WC. et. al. (2017) proposed a method based on Siamese CNNs approach and showed how to use the proposed GAN model for outfit recommendations [13]. The model takes a text query and a history of user outfits and generates personalized fashion recommendations. The authors compared their method with some baselines and showed that it outperforms such base methods as WARP, FM, BPR (and some variants), and others.

Sudhir K. and Mithun D.G. (2019) combined the encoder-decoder architecture with GAN approach and proposed the model which takes a vector of features of input item image and random noise to generate a new item [12]. The noise is used to diversify the generated items.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

Yu C. et. al. (2019) also used encoder-decoder-based GAN to explore compatibility of items and generate compatible items for outfits [16]. An encoder-decoder-based generator was used to generate an item by an input one. Two same discriminators were used to compute the compatibility score of the built outfit and evaluate whether a real item was generated or not. The authors proposed to use the BPR-based method to compute compatibility by ranking a positive, a generated, a negative and a random compatible outfit. They showed that their method is more accurate than directly assessing by classifying or scoring the generated outfit.

Liu L. et. al. (2019) proposed the Attribute-GAN model for clothing matching [14]. They added the second discriminator to assess attributes of synthetic and negative images. To extract attributes from the synthetic image, the model projects it to a vector of visual features, splits the vector into parts, and uses several dense layers to project them to attributes. Extracted one-hot encoded attributes and attributes of the negative item are used to calculate the loss. The real-fake discriminator is used to calculate the second part of the loss function.

Liu J. et. al. (2020) proposed the multimodal method to generate an item by an input one [15]. It combines the encoder-decoder-based GAN and TextCNN to generate a new item by an input item. The encoder projects an input image to a vector of visual features as TextCNN projects an input description to a vector of context features. Vectors of features are concatenated and used by the decoder to generate a new item. The authors used a loss function that combines four parts: BPR loss, pixel difference between generated image and corresponding ground truth image from a dataset, compatibility of input and ground truth images, and compatibility of their descriptions.

The first problem with all of the GAN-based methods is that they use a fixed number of input and output items in the set. The models can only operate with the number of items specified during the training, and changing this value will require to retrain the model. The second problem is that the existing solutions either use only the visual features and extract content attributes from generated images, or process text separately from generation. However, such approaches do not guarantee keeping the attributes of the input items and the reliability of the evaluation of the attributes of the received items. We are trying to solve both problems mentioned above in this study.

3 Transformer-based GAN for outfit recommendations

In this section, we describe the proposed method that tries to improve the compatibility of set items, to solve the problem of a fixed number of input and output items and to decrease the content features vanishing during generation. First, we define the problem and introduce the method of representing a set of items as a sequence of items. Then we describe our multimodal transformer block with cross-modal attention for simultaneous exploration of the compatibility of items inside a modality and between modalities. Finally, we present our multimodal transformer-based GAN with cross-modal attention.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

3.1 Problem formulation

Suppose we have some item domains $D = \{D_1, \dots, D_N\}$ and a domain S of sets of items with a variable number of items. Each item $I_{i,j} \in D_i, i \in 1 \dots N, j \in 1 \dots |D_i|$ is associated with a visual image $Vis_{I_{i,j}}$ and a textual description $C_{I_{i,j}}$ (C as content). Since sets of items have a variable number of items, we extend them to a pre-defined maximum number of items M with a random normal noise $N = (0, 1)$ as placeholders, then an extended set of items $S_k = \left\{\left(Vis_{I_1^k}, C_{I_1^k}\right), \left(Vis_{I_2^k}, C_{I_2^k}\right), \dots, \left(Vis_{I_n^k}, C_{I_n^k}\right)\right\}$, $|S_k| = n$ can be described as $\hat{S}_k = \{S_k, \left(Vis_{N_{n+1}^k}, C_{N_{n+1}^k}\right), \dots, \left(Vis_{N_M^k}, C_{N_M^k}\right)\}$, where Vis_N^k, C_N^k are randomly sampled vectors of visual and textual features. We focused on devising an end-to-end multimodal generative compatibility modeling scheme Sch that is able to learn the compatibility c between a set of items and project an input noise to synthetic items $\tilde{I}$, by introducing the network G as follows:

$$\begin{gathered} G\left(\hat{S}_k \mid \Theta_{\mathrm{G}}\right) \rightarrow \tilde{S}_k; \\ \tilde{S}_k = \left\{S_k, \left(Vis_{\tilde{I}_{n+1}^k}, C_{\tilde{I}_{n+1}^k}\right), \dots, \left(V_{\tilde{I}_M^k}, C_{\tilde{I}_M^k}\right)\right\}; \\ c_k = Sch\left(\tilde{S}_k \mid \Theta^{\mathrm{Sch}}\right), \end{gathered} \tag{1}$$

where Θ^G and Θ_{Sch} are a set of parameters to be learned of generator G and scheme Sch, c_k – is a compatibility score of generated set $\tilde{S}_k$.

3.2 Multimodal transformer-based GAN with cross-modal attention

Multimodal transformer block. As the exploring of compatibility of items and the generation of new compatible items are the main tasks of our scheme, we can use the self-attention module, which is already presented in traditional transformer architecture [24]. It allows exploring the compatibility of items inside a single modality, for example, a visual or textual modality. However, each item is associated with data of multiple modalities. The usage of multi-way scheme, which explores each modality separately and then fuzes them, is not an optimal solution because the final vector contains features of multiple modalities but features of each modality are explored without others and are not connected with them.

To explore compatibility of items between the modalities, we propose to use a cross-modal attention (CMA) module firstly described in Click or tap here to enter text.. Similar to the classical self-attention module, each sequence of items is represented as query (Q), key (K), value (V), however, K and V are swapped between modalities: $K_{Vis} \rightarrow K_C, K_C \rightarrow K_{Vis}, V_{Vis} \rightarrow V_C, V_C \rightarrow V_{Vis}$, where K_{Vis} and K_C are keys of a visual and textual modalities correspondingly, V_{Vis} and V_C are values of a visual and textual modalities correspondingly.

To exploit the advantages of self-attention and cross-modal attention modules simultaneously, we propose to stack these modules as shown in Fig. 1. First, we propose to forward vectors to the cross-modal attention module and obtain vectors of features that contain compatibility information between items of the same modality and between modalities.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

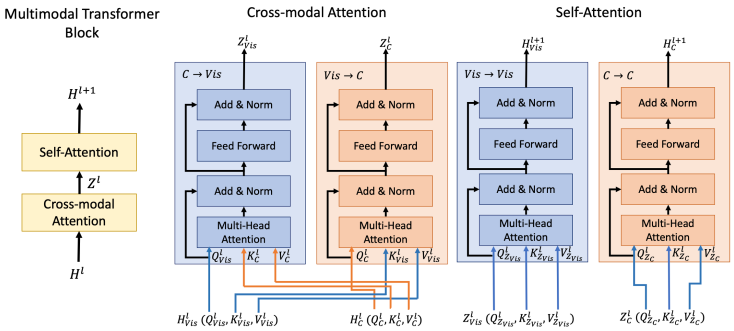

Fig. 1. Multimodal Transformer Block architecture (left) with cross-modal attention (center) and self-attention (right) modules.

Fig. 1. shows the multimodal transformer block, which explores two modalities: a visual and a textual, but it can be easy to extend the scheme with additional modalities.

Multimodal transformer-based GAN. The problem of fitting a model to recommend items combined with input ones before is important in our opinion. It severely limits a variety of compatible items and, as a result, a variety and a number of sets. The obvious and efficient solution to smooth this problem is to use a GAN [12–16]. We propose to stack several multimodal transformer blocks described above and use them as a body of a generator of our GAN-based model, as shown in Fig. 3.

The generator G aims to translate the given sequence of items $\hat{S}_t$ aligned between modalities, which contains a noise as a placeholder to the missed items, to a compatible with non-noise items sequence of items $\hat{S}'_t$ with compatibility score c_t. The generator also aims to predict a variable number of items in a target sequence.

The standard method to learn the GAN-based generator is to use a real-fake discriminator. But, in fact, the traditional real-fake discriminator can only enforce the generator to produce realistic vectors. These vectors can be incompatible with input ones. In our context, we need not only to synthesize realistic vectors but also learn the compatibility of input vectors and synthesize new compatible vectors with them. To achieve this, we propose to use (in addition to a real-fake discriminator $Discr_{rf}$) a compatibility discriminator $Discr_{comp}$ as the guidance for compatibility modeling.

The body of discriminators is similar to the generator body. The real-fake discriminator takes a target sequence of items and projects it to a sequence of latent representations. Then, dense layer is used for each item of sequence to compute pre-item real-fake scores. Finally, real-fake scores are summed to compute a complete real-fake score and corresponding real-fake loss.

To overcome the problems of traditional GANs, [16,26] proposed to use a combination of "relativistic discriminator" [26] and LSGAN [27] and described the following in Equation 2 losses for the real-fake discriminator and generator.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

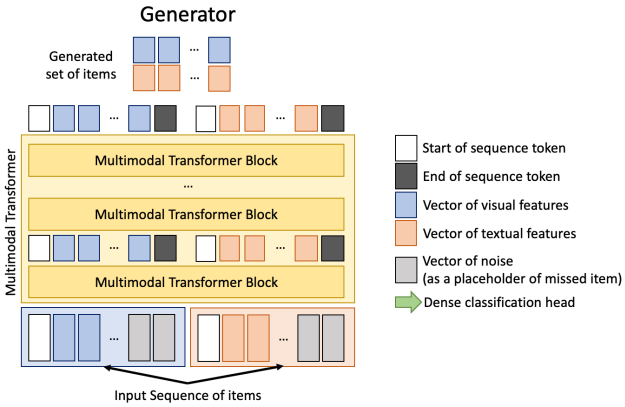

Fig. 2. Scheme of Multimodal Transformer. It is a base of Generator and Discriminators.

$$L_{rf}^{Discr} = \frac{1}{2}\mathbb{E}_{o^r}\left[\left(s_{rf}(o^r) - \mathbb{E}_{o^f}s_{rf}(o^f) - 1\right)^2\right] + \frac{1}{2}\mathbb{E}_{o^f}\left[\left(s_{rf}(o^f) - \mathbb{E}_{o^r}s_{rf}(o^r) + 1\right)^2\right] \mid \Theta_{rf}^{Discr}, \tag{2}$$

where s_{rf} is a real-fake score, $o^r \sim O^r$, $o^f \sim O^*$, O^r is a domain of real sets, O^* is a domain of synthesized sets of items, Θ_{rf}^{Discr} is a set of parameters to be learned of the real-fake discriminator. It can be seen that the discriminator keeps a margin between real and fake data. The generator should eliminate this gap by minimizing:

$$L_{rf}^{G} = \frac{1}{2}\mathbb{E}_{o^r}\left[\left(s_{rf}(o^r) - \mathbb{E}_{o^f}s_{rf}(o^f)\right)^2\right] + \frac{1}{2}\mathbb{E}_{o^f}\left[\left(s_{rf}(o^f) - \mathbb{E}_{o^r}s_{rf}(o^r)\right)^2\right] \mid \Theta^G, \tag{3}$$

where Θ^G is a set of parameters to be learned of the generator. As [16] uses a single item as input, we re-defined the real-fake score function as a function to compute the real-fake score of a set. It calculates a score of a set as a summation of a per-item real-fake score of each item inside each modality as follows:

$$s_{rf}(\tilde{S}_k) = \sum_{i=0}^{|\tilde{S}_k|} s_{rf}(I_i^k) = \sum_{j=0}^{|\tilde{S}_k|}\left(s_{rf}(Vis_i^k) + s_{rf}(C_i^k)\right) \tag{4}$$

The compatibility discriminator similarly takes a target sequence and projects it into latent space. To obtain the compatibility score of a set of items by the latent representation, we have modified the scheme of computing the compatibility score by ranking, which is proposed in Click or tap here to enter text.. We first take the element-wise product of each pair of items inside each modality and sum them to obtain a latent space representation z of the set:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

$$z(\tilde{S}_k) = \sum_{i=0}^{|\tilde{S}_k|} \sum_{j=0}^{|\tilde{S}_k|} I_i^k \odot I_j^k, i \neq j;\ I_i^k \odot I_j^k = \left(Vis_i^k \odot Vis_j^k + C_i^k \odot C_j^k\right) \quad (5)$$

Then we fed the result into a metric network M, which consists of several dense layers, to get the final compatibility score $s_{comp}(\tilde{S}_k) = M(z(\tilde{S}_k) \mid \Theta_M)$, where Θ_M is a set of parameters to be learned of metric network M, s_{comp} is a compatibility score.

To train the compatibility discriminator, we split our dataset into positive O^+ and negative O^- sets as described in [16]. The compatibility discriminator should be able to distinguish positive sets from negative ones by assigning higher compatibility scores to positives $s_{comp}(O^+) > s_{comp}(O^-)$.

To achieve this, the compatibility discriminator should seek to reduce the loss:

$$L_{comp}^{Discr} = -\mathbb{E}_{\substack{o^+ \sim O^+ \\ o^- \sim O^-}} \left[ln\left[\sigma\left(s_{comp}(o^+) - s_{comp}(o^-)\right)\right]\right] + \lambda_{\Theta_M} \mid \Theta_{comp}^{Discr}, \quad (6)$$

where σ is the sigmoid function, λ is a regularization term, Θ_{rf}^{Discr} is a set of parameters to be learned of the compatibility discriminator, $\Theta_M \in \Theta_{rf}^{Discr}$, $o^+ \sim O^+$, $o^- \sim O^-$. To achieve this by generator, it should synthesize a set $\tilde{S}_k \sim O^*$ with a similar compatibility score as its positive set S_k^+. As a result, it should seek to reduce the loss:

$$\begin{aligned} L_{comp}^G = &\frac{1}{2}\mathbb{E}_{o^+}\left[s_{comp}(o^+) - \mathbb{E}_{o^*}\left(s_{comp}(o^*)\right)^2\right] \\ &+ \frac{1}{2}\mathbb{E}_{o^*}\left[s_{comp}(o^*) - \mathbb{E}_{o^+}\left(s_{comp}(o^+)\right)^2\right] \mid \Theta^G, \end{aligned} \quad (7)$$

where $o^* \sim O^*$.

The overall architectures of the real-fake and the compatibility discriminators are shown in Fig. 3.

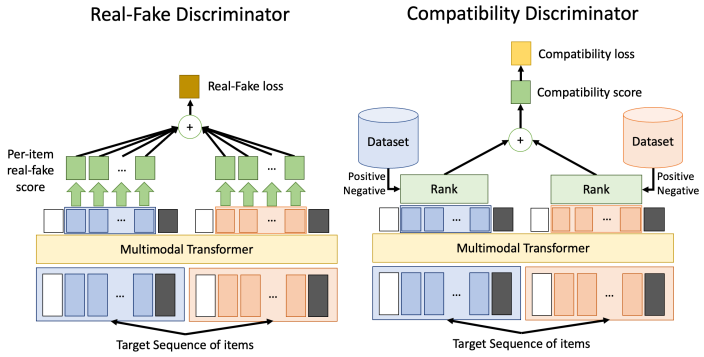

Fig. 3. Scheme of Real-Fake Discriminator (left) and Compatibility Discriminator (right).

It should be noted that the multimodal transformers of discriminators have shared parameters $\Theta_{shared}^{Discr}, \Theta_{shared}^{Discr} \in \Theta_{rf}^{Discr}, \Theta_{shared}^{Discr} \in \Theta_{comp}^{Discr}$.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

The final objective of our generator is to minimize loss function as follows in Equation 10, and the final objective of our complete scheme is to minimize loss function as follows in Equation 11:

$$L^G = \lambda_1 L^G_{rf} + \lambda_2 L^G_{comp}; \tag{8}$$

$$L = L^G(\lambda_1, \lambda_2) + \lambda_3 L^{Discr}_{rf} + \lambda_4 L^{Discr}_{comp}, \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are model tradeoff parameters.

The training process can be described as follows. First, a step is made on the discriminators, after which they are updated to estimate the generator. A result of forwarding a batch to the generator is used to calculate losses of discriminators, but the losses are backwarded only to corresponding discriminators. Then the step is made on the generator with estimates from the discriminators. Similar to discriminators, the loss is backwarded only to the generator. This process continues until the generator converges.

4 Experiments

4.1 Dataset

Most of the previous works for outfit recommendation have used either Amazon data containing co-purchase information or Polyvore data containing outfits created by users. Co-purchase does not always mean that items are compatible as items are typically not bought with the intention of being worn together, but it is more likely to reflect a user's style preference [9,23]. As a result, the estimate on this dataset is not reliable. Outfits in datasets obtained from Polyvore are built by users which gives a stronger signal of compatibility. They contain a variable number of items per outfit, visual and textual modalities, and some other data. They are fully suitable for evaluation on them in terms of the content, but they are outdated. Items from them are out of fashion and mostly not available. We decided to use the public dataset FOTOS which contains outfits and corresponding items [19,20]. Each outfit and each item are associated with an image and metadata. Outfits contain a variable number of items, and they are created by users, similar to Polyvore. It consists of 10,988 compatible outfits and 20,318 items.

4.2 Baselines

To verify the effectiveness of proposed method, we compared it with the following baseline methods. **FHN** uses only visual features. It encodes them with category encoders and then learns one-hot encodings for item embeddings. The outfit score is the mean of pairwise compatibility scores of outfit items. **Bi-LSTM-VSE** is interpreting an outfit as a sequence of items and exploits the outfit compatibility by a bi-directional LSTM and visual-semantic consistency. **NGNN** is a node-wise graph-based neural network that optimizes a multimodal fashion graph to uncover the complex relationships among items and assess the compatibility score. **HFGN** is a hierarchical graph neural network to model relationships among users, items and outfits simultaneously. It uses message propagation across items and attention to better capture compatibility between

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

items. The model contains two levels for exploring interactions between users and outfits, and between outfits and items correspondently. For our experiment, we have used only the second one. **MGCM** is an autoencoder-based GAN model that uses deep CNN to extract visual features from images and TextCNN to extract textual features from descriptions. It explores the visual and textual features simultaneously.

It is worth noting some implementations details. MGCM generates images by default, but this is not necessary for this task. The model has been adapted to generating feature vectors instead of images. It also takes one item as input and generates one item as output. To generate one item by multiple ones, the encoded vectors of items have been averaged and the obtained vector have been used as input for the generator. MGCM, and our transformer-based GAN model are generating synthetic vectors of features. To assess the accuracy of models, we are comparing generated vector with vectors of candidates and choosing the closest one.

4.3 Evaluation and results

The proposed method and baselines were compared on the fill-in-the-blank task. For each outfit in the dataset, a random item was selected as the blank. Similarly, three negative candidates were randomly selected for each outfit. The aim is to select the correct answer from four candidates to fill in the blank in the outfit. The accuracy of assessing the performance was proposed in Table 1. The best result is in bold, the second score is underlined, and the third score is in italic.

Table 1. Performance comparison on fill-in-the-blank task.

Method	FLTB (2 items)	FLTB (3 items)	FLTB (4 items)
FHN	0.697	0.669	0.669
Bi-LSTM-VSE	0.724	0.776	0.753
NGNN	0.791	0.765	0.765
HFGN	0.826	*0.801*	*0.800*
MGCM	*0.822*	0.817	0.829
Our	**0.878**	**0.863**	**0.911**

The results show that our proposed model outperforms the baselines. They also illustrate that GAN-based methods outperform the classical ones, despite the fact that GAN learn to generate new items rather than choose from existing ones.

Fig. 4. and Fig. 5. show some examples where our model predicts blanked items better than others compared. Fashion and style are very complex concepts, and preferences can differ from person to person. In our opinion, the figures show that HGFN and MGCM models emphasized various aspects of the input items, while our model evaluated the overall style of the received items and their attributes and generated the most appropriate item for the set.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

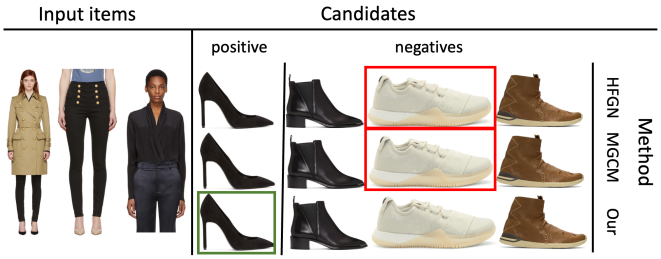

Fig. 4. Example of fill-in-the-blank predictions of blanked "shoes" item by input "outer", "bottom" and "top" items.

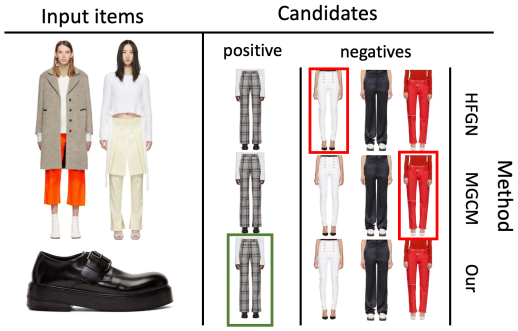

Fig. 5. Example of fill-in-the-blank predictions of blanked "bottom" item by input "outer", "top" and "shoes" items

5 Conclusions

In this paper, we propose transform-based GAN methods for multimodal generation of set of items by input ones. We have tried to solve some problems of existing GAN-based solutions such as the problem of a fixed number of input and output items and the problem of losing content attribute features during the generation process.

We have proposed to interpret a set of items as a sequence of items with starting and ending tag. It allows a model to control a number of items which a set of items contains.

The multimodal transformer with cross-attention module have been proposed as a body of our GAN-based model to overcome the problem of losing a content attribute features during generation. It allows the model to explore and generate visual and textual features simultaneously.

We have compared the proposed model with some strong baseline models and shown that our proposed method outperforms baselines at the fill-in-the-blank task on the FOTOS dataset. Our model has reached 0.878 accuracy in filling one blanked item by one of four candidates (one is positive and three are negatives) by one input item. Versus 0.724 of Bi-LSTM-VSE, 0.822 of MGCM, 0.826 of HFGN. Similarly, our model outperforms other baseline models with a different number of input items.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement №17-71-30029 with co-financing of Bank Saint Petersburg.

References

1. Han X, Wu Z, Jiang YG, Davis LS. Learning Fashion Compatibility with Bidirectional LSTMs. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*. July 2017:1078-1086. doi:10.1145/3123266.3123394
2. Song X, Feng F, Liu J, Li Z, Nie L, Ma J. NeuroStylist: Neural Compatibility Modeling for Clothing Matching. *Proceedings of the 25th ACM international conference on Multimedia*. 2017:753-761. doi:10.1145/3123266.3123314
3. Yang X, Ma Y, Liao L, Wang M, Chua TS. TransNFCM: Translation-Based Neural Fashion Compatibility Modeling. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. December 2018:403-410. doi:10.1609/aaai.v33i01.3301403
4. Tangseng P, Yamaguchi K, Okatani T. Recommending Outfits from Personal Closet. *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2017:2275-2279.
5. Lu Z, Hu Y, Chen Y, Zeng B. Personalized Outfit Recommendation with Learnable Anchors. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021:12722-12731.
6. Li X, Wang X, He X, Chen L, Xiao J, Chua TS. Hierarchical Fashion Graph Network for Personalized Outfit Recommendation. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. May 2020:159-168. doi:10.1145/3397271.3401080
7. Lu Z, Hu Y, Jiang Y, Chen Y, Zeng B. Learning binary code for personalized fashion recommendation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2019;2019-June:10554-10562. doi:10.1109/CVPR.2019.01081
8. Cui Z, Li Z, Wu S, Zhang X, Wang L. Dressing as a Whole: Outfit Compatibility Learning Based on Node-wise Graph Neural Networks. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. February 2019:307-317. doi:10.1145/3308558.3313444
9. Cardoso A, Daolio F, Vargas S. Product Characterisation towards Personalisation: Learning Attributes from Unstructured Data to Recommend Fashion Products. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. March 2018:80-89. doi:10.1145/3219819.3219888
10. Sagar D, Garg J, Kansal P, Bhalla S, Shah RR, Yu Y. PAI-BPR: Personalized Outfit Recommendation Scheme with Attribute-wise Interpretability. *Proceedings - 2020 IEEE 6th International Conference on Multimedia Big Data, BigMM 2020*. August 2020:221-230. doi:10.1109/BigMM50055.2020.00039
11. Yuan F, Karatzoglou A, Arapakis I, Jose JM, He X. A Simple Convolutional Generative Network for Next Item Recommendation. *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. 2018;19:582-590. doi:10.1145/3289600.3290975
12. Kumar S, Gupta M das. c^+GAN: Complementary Fashion Item Recommendation. *KDD'19: Workshop on AI for Fashion*. June 2019. https://arxiv.org/abs/1906.05596v1. Accessed January 21, 2022.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12

13. Kang WC, Fang C, Wang Z, McAuley J. Visually-Aware Fashion Recommendation and Design with Generative Image Models. *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2017;2017-November:207-216. doi:10.1109/ICDM.2017.30
14. Liu L, Zhang H, Ji Y, Jonathan Wu QM. Toward AI fashion design: An Attribute-GAN model for clothing match. *Neurocomputing*. 2019;341:156-167. doi:10.1016/J.NEUCOM.2019.03.011
15. Liu J, Song X, Chen Z, Ma J. MGCM: Multi-modal generative compatibility modeling for clothing matching. *Neurocomputing*. 2020;414:215-224. doi:10.1016/J.NEUCOM.2020.06.033
16. Yu C, Hu Y, Chen Y, Zeng B. Personalized Fashion Design. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019;2019-October:9045-9054. doi:10.1109/ICCV.2019.00914
17. Hsiao WL, Grauman K. Creating Capsule Wardrobes from Fashion Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. December 2017:7161-7170. doi:10.1109/CVPR.2018.00748
18. Dong X, Jing P, Song X, Xu XS, Feng F, Nie L. Personalized capsule wardrobe creation with garment and user modeling. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*. October 2019:302-310. doi:10.1145/3343031.3350905
19. Zheng N, Song X, Niu Q, Dong X, Zhan Y, Nie L. Collocation and Try-on Network: Whether an Outfit is Compatible. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*. October 2021:309-317. doi:10.1145/3474085.3475691
20. Dong X, Wu J, Song X, Dai H, Nie L. Fashion Compatibility Modeling through a Multi-modal Try-on-guided Scheme. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. July 2020:771-780. doi:10.1145/3397271.3401047
21. Vasileva MI, Plummer BA, Dusad K, Rajpal S, Kumar R, Forsyth D. Learning Type-Aware Embeddings for Fashion Compatibility. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2018;11220 LNCS:405-421. doi:10.1007/978-3-030-01270-0_24
22. Han X, Wu Z, Jiang YG, Davis LS. Learning Fashion Compatibility with Bidirectional LSTMs. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*. July 2017:1078-1086. doi:10.1145/3123266.3123394
23. Bettaney EM, Hardwick SR, Zisimopoulos O, Chamberlain BP. Fashion Outfit Generation for E-commerce. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2019;12461 LNAI:339-354. doi:10.1007/978-3-030-67670-4_21
24. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*. 2017;2017-December:5999-6009. https://arxiv.org/abs/1706.03762v5. Accessed January 21, 2022.
25. Cheng Y, Wang R, Pan Z, Feng R, Zhang Y. Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. 2020;20:3884-3892. doi:10.1145/3394171.3413869
26. Jolicoeur-Martineau A. The relativistic discriminator: a key element missing from standard GAN. *7th International Conference on Learning Representations, ICLR 2019*. July 2018. https://arxiv.org/abs/1807.00734v3. Accessed January 21, 2022.
27. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least Squares Generative Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*. 2016;2017-October:2813-2821. doi:10.1109/ICCV.2017.304

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_12