

# Multimodal Representation Learning for Fashion-Item Recommendation System

Sojeong Song

akqjq4985@kaist.ac.kr

## Abstract

*The dominant multi-modal recommendations are based on user's behavior information such as product purchase history, rating, and review. We present a new multi-modal recommendation approach that uses conversation data between a user and a coordinator. Our goal is to extract universal features in the fashion domain (e.g., product classification, keyword tagging, outfit recommendation) from fashion images and descriptive text information, and use these features to recommend appropriate fashion items according to the user's utterances. To achieve the goal, we followed three steps: (1) preprocess and refine the FASCODE dataset; (2) rectify our recommendation model and evaluation algorithms; and (3) perform performance improvement research. Our model makes joint embedding space for mapping item images and descriptions through the Projection network and transfers the information of dialog to the item embedding space through the Transfer network. It is proved that our model (1) improves the performance by utilizing richer information, (2) keeps the performance when some text information is lost, and (3) is efficient. However, there are some limitations such as low WKT value and weak fusion network. Therefore, future research should be conducted in adding some techniques such as contrastive loss or generative adversarial network to predict more accurate coordi-sets.*

## 1. Introduction

Multimodal recommendation systems typically include collaborative filtering and content-based filtering. Collaborative filtering provides recommendations based on user behaviors, such as ratings or product purchase history. One of these methods is Nearest neighbor collaborative filtering, which predicts the user's preference for the item based on the user-item matrix. Another method is latent factor-based collaborative filtering, which extracts latent factors by reducing dimensionality through matrix factorization. It has the advantage of high memory efficiency. Examples of this method include MF [11], NGCF [14], and LightGCN [6]. On the other hand, content-based filtering is to recommend

items similar to those preferred by users. Example of this method include VBPR [5], MMGCN [15], and GRCN [17]. Among them, we decided to generate a model based on the VBPR.

VBPR first uses features extracted from item images with item latent factors. It shows that using visual features improves the existing model's performance by about 10% and solves the cold start issues, which make it difficult to recommend a new item because there is no purchase history. This paper concludes that items with similar styles are located close to each other in the 2D visualization of the embedding space. It implies that the embedding networks learn hidden taxonomy to map and users correctly.

Most recommendation systems are based on statistical user's behaviors, but they do not focus on how to fit the user's specific situation. This is because it is difficult to grasp the user's needs according to the situation due to the absence of direct interaction with the user. The most intuitive way to catch the user's intention is through conversion, however, the VBPR is inappropriate to use image and text together and is only limited to the image dataset. For this reason, we present a novel approach to a recommendation system that recommends fashion items through conversations with users. We aim to train a model that recommends accurate fashion coordi-sets by using meta-data of fashion items and dialog between user and coordinator. Therefore, we need to create a new model to integrate the item features and dialog boxes due to no existing prior work.

To address this, we will train the ProjectionNet to project image and text embedding vectors to the target dimension and learn the mapping between two embedding spaces. Then, the information of dialog is extracted and transferred to the item embedding space through the TransferNet. In the evaluation process, we will measure the Mean Square Difference similarity between dialog and whole items at the target dimension, which is called the reduced dimension in this paper, and pick four items that have the most similar to the dialog for predictions of each type. These predictions are used to rank the candidates in the evaluation dataset. We implement our model and perform many experiments to improve the performance, and the results show that our model learns the joint embedding space well to exploit im-

age and text information together, so the performance does not drop when some of the text information is lost. These are our main contributions and the details will be described in further sections:

- Preprocess and refine the raw text files to dictionaries of embeddings.
- Recify our own baseline and evaluation algorithm.
- Performance improvement research figuring out the effects of hyperparameters and neural architectures.

## 2. Related Work

### 2.1. Multimodal machine learning

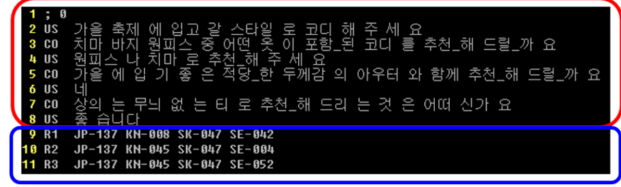
Multimodal machine learning is to learn from multimodal sources and expects the possibility of capturing correspondences between modalities gaining an in-depth understanding of natural phenomena. The field of study of multimodal learning presents several challenges when considering the heterogeneity of data. There are five following challenges: Representation, Translation, Alignment, Fusion, and Co-learning [1]. Among these challenges, our work is the task of multimodal representations, which learn how to represent and summarize multimodal data. It is important to utilize the complementarity and redundancy of multiple modalities. The multimodal representation can be divided into two categories, joint, and coordinated [1].

Joint representations project unimodal signals together into a multimodal space. It is usually used where multimodal data is in both during the training and inference process. For example work, Ngiam et al. [9] proposed the joint representation model that uses autoencoders in the multimodal domain. In addition, Srivastava [13] introduced multimodal deep belief networks as a joint representation. Furthermore, the model presented by Kim et al. [7] uses a deep belief network before combining each modality.

Coordinated representation deal separately with unimodal representations, but coordinate them through constraints. One of the examples of this method is work by Weston et al. [16], who proposed the WSABIE model. Its coordinated space was implemented for images and their annotations. In addition, Kiros et al. [8] coordinate the feature space by extending WSABIE to sentence and image coordination representations through the LSTM model.

### 2.2. Pretrained models for processing image and text

For embedding process, we will use pretrained model KoELECTRA [10] and Img2vec [12]. KoELECTRA is the ELECTRA model that is trained with Korean text. We chose this since it can be used on any OS and does not require a tokenization file while ensuring good performance.



1	:	0
2	US	가을 축제 에 입고 갈 스타일 로 코디 해 주 세 요
3	CO	치마 바지 원피스 중 어떤 것 이 포함 된 코디 를 추천 해 드릴 까 요
4	US	원피스 나 치마 로 추천 해 주 세 요
5	CO	가을 에 입 기 좋 은 적당 한 두께감 의 아우터 와 함께 추천 해 드릴 까 요
6	US	네
7	CO	상의 는 무늬 없 는 티 로 추천 해 드 리 는 것 은 어 떠 신 가 요
8	US	중 습 니 다
9	R1	JP-137 KN-008 SK-047 SE-042
10	R2	JP-137 KN-045 SK-047 SE-044
11	R3	JP-137 KN-045 SK-047 SE-052

Figure 1. Example of evaluation dataset. The red box represents the dialog between the user and the coordinator. The blue box contains a set of candidate coordi-sets.

Img2Vec is Python library for extracting vector embeddings for any image and shows good performance on various tasks such as ranking for recommender systems, clustering images to different categories, classification tasks, and image compression.

## 3. Method

### 3.1. FASCODE dataset

FASCODE(FASHion COordination DatasEt / FASHion CODE) [2] is a fashion dataset with meta-data of fashion-item and a dialog dataset between an AI fashion coordinator and a user. Meta-data of fashion items includes a total of 2603 fashion items, and they are labeled according to what kind of clothes they are. The items are described in four aspects: shape, material, color, and mood (Tab. 1a). And also, each of them is assigned to appropriate item classes (Tab. 1b). And a dialog dataset includes information about a speaker, the order of the dialog, tagging for each dialog (Tab. 2). Note that fashion-item descriptions and dialog are written in Korean.

The dataset for evaluation consists of sets of dialog and its candidate coordi-sets. The challenge is to look at the conversation and find the exact priority of the candidate coordi-set recommended to the user. Figure 1 shows an example evaluation data. There are three candidates and the answer in the blue box. The task is to use only the dialogues in the red box to find the priority for the three coordinate sets in the blue box.

### 3.2. Embedding process

In embedding process, we obtain three embeddings of dialog, item image, and item text by passing dialog. Figure 2 shows embedding process briefly. KoELECTRA is used for text modality, and Image2Vec with Resnet-18 is used for the image. We will train with recommended items for each dialog in the training dataset.

### 3.3. Training process

Figure 3 shows training process. We use projection network and transfer network in training process. Transfer network is composed of 4 blocks, each block is composed of

Feature	Code
Shape	F
Material	M
Color	C
Mood	E

(a) Fashion item description.

Class	Code	Item
Outwear	O	Jacket, Jumper, Coat, Cardigan, Vest
Top	T	Knitwear, Sweater, Shirt, Blouse
Bottom	B	Skirt, Pants, Dress
Shoes	S	Shoes

(b) Fashion item class.

Table 1. Fashion item meta-data in FASCODE.

Speaker	Conversation & Tagging
<US>	Please recommend a trench coat.
<AC>	CT-019
<CO>	This is a trench coat that goes well with your inner color. EXP_RES_TYPE;EXP_RES_COLOR
<US>	Please also recommend shoes. USER_SUCCESS
<CO>	Do you prefer sneakers or high heels? ASK_TYPE
<US>	Please recommend sneakers.
<AC>	SE-039
<CO>	This is a basic item that goes well with any style. EXP_RES_ETC
<US>	I like it. Can I see the full coordination? USER_SUCCESS
<AC>	CT-019 SW-009 SK-053 SE-039

Table 2. Example dialog in train dataset of FASCODE. Note that this is translated dialog.

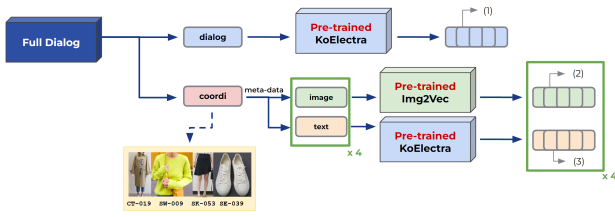


Figure 2. Embedding process: (1) embeddings of dialog (2) embeddings of item image (3) embeddings of item description.

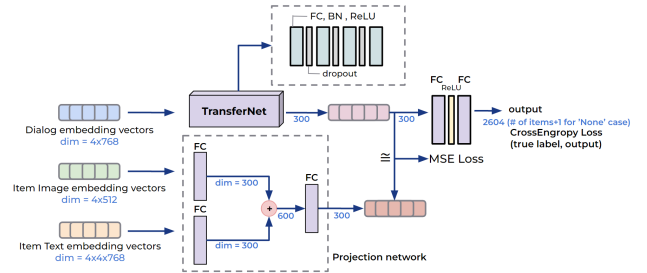


Figure 3. Training process. The transfer network is used for dialog embedding vectors, and the projection network is used for image embedding vectors and item text embedding vectors.

a fully connected layer, batch normalization, and ReLU activation function, and a dropout layer is inserted between them. The dialog embedding vectors transmit the information to the item embedding space by passing through the transfer network. And through the projection network, the item embedding space is created by combining image embedding vectors and item text embedding vectors and passing through a fusion layer. Each embedding vector is projected to the same dimension as each other and trained to affect each other equally. The output of transfer network and projection network has the same dimension that is called reduced dimension. We use MSE loss so that the two vectors are similar to each other. The output of the transfer network is trained using cross entropy loss to predict item id through two layers.

### 3.4. Evaluation process

In the evaluation process, the dialog included in the evaluation dataset and the three candidate coordi-sets corresponding to each dialog are made into embeddings using pretrained models, respectively. Figure 4 shows the evaluation process. All three embeddings are turned into comparable vectors by passing through a well-trained network. And every fashion item goes through the projection network to make it comparable.

The detailed evaluation method is described in Figure 5. We compare the embedding vectors extracted from the dialog with the 2604 embedding vectors from the entire item and none case, then the four most similar for each type are extracted. For comparison, we produce the predicted coordi-sets by choosing one with the smallest MSE. And by

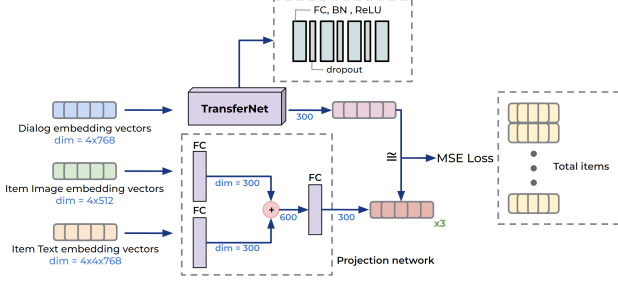


Figure 4. Evaluation process. Dialog and item embedding vectors pass through a well-trained transfer and projection network, respectively.

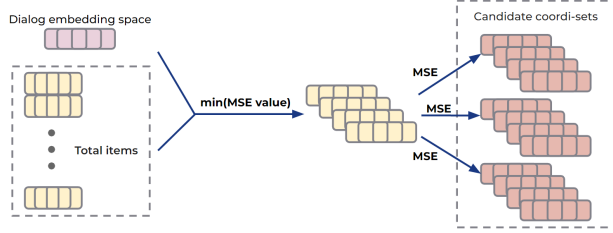


Figure 5. Evaluation method. Extract four most similar embedding vectors for each type, and measure the ratings using WKT.

measuring MSE with vectors selected from candidates for each type, the coordi-set with the smallest MSE becomes rating 1, and the coordi-set with the highest MSE becomes rating 3 to predict the rating. These predicted ratings are measured by Weighted Kendall’s Tau compared to the actual rating, which is a kind of rank correlation coefficient. It measures weighted correlation coefficients by comparing the rank of two variables. Higher scores means better recommendation results.

## 4. Experiment

### 4.1. Experimental setup

We use Pytorch to implement our model and 4 NVIDIA DGX A100 GPUs are used for the experiments. We train the network using SGD optimizer with a momentum of 0.9 and learning rate of 0.01. The dropout ratio is 0.2 and the batch size is 64. The embedding vector dimension of the image is 512 and the text is  $768 \times 4$  or  $768 \times 16$ . The reduced dimension is 150, 300, or 600, and 300 is a default value. The probability of missing information is 0.15.

### 4.2. Compared methods

We use four compared methods for evaluation. (1) Experiments about the effects of using different modalities. We compare the case of using only text and the case of using images and text together. We can see how using multiple modalities affects learning. (2) Missing some information

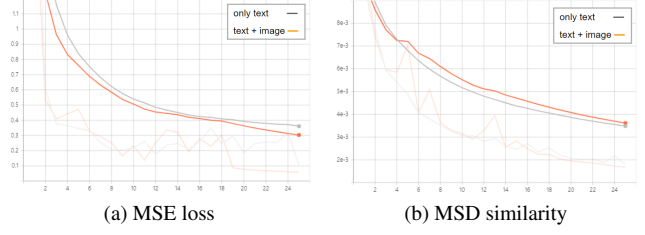


Figure 6. Loss and similarity results of text-only model and text+image model.

of target modality. We test the degree of performance degradation on each text-only and text+image model when some text or image information data is lost during the training process. We can confirm how well the multimodal representation learning is well trained. (3) Change the neural networks in the baseline by varying the number of layers and reduced dimensions. We expect to get an insight into performance improvement.

## 5. Experimental Results & Analysis

### 5.1. Effects of using different modality

The comparison result of text-only and text+image model is represented in Figures 6a and 6b. In this experiment, the embedding dimension of KoELECTRA model is set to  $768 \times 16$ . These show each MSE loss and Mean Square Difference(MSD) similarity over training steps. Both showed similar performance, but the text-image model showed slightly better results. The comparable performance of multimodal model is very common in other models of multimodal representation learning. From these results, we conclude that the text+image model can predict more accurately because it uses richer information successively through multiple modality.

### 5.2. Missing information of target modality

**Missing text information** To confirm the performance of our multimodal network, we try to observe the degree of performance degradation after removed 15% of the text information. Figures 7a and 7b shows the results of this experiment. As a result, there is not much performance drop even when a large parts of text information is missing. This is because the lost text information is supplemented with image information, which proves that our model is very robust to noise in the item text dataset.

**Missing image information** We also try to remove 15% of image information at training time by tracking the MSE loss and MSD similarity. Figures 7 and 7c shows the results, and it shows lower loss and similarity when some of image information is lost. We think this is because the 15% of image embedding vectors are changed to zero vectors, so the learning becomes easier. The results of WKT is 0.1409

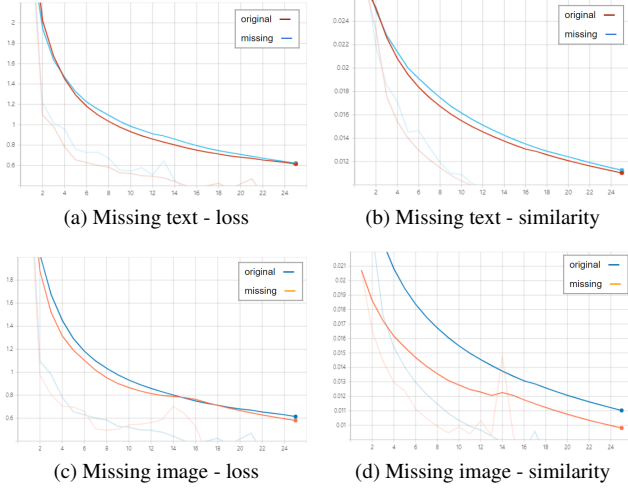


Figure 7. Loss and similarity results of missing information of target modality experiment. (a)(b) Missing text (c)(d) Missing image.

with whole data and 0.1209 with 75% image information. It can be observed that there are little degradation of WKT values when some images are lost.

### 5.3. Change the neural networks in baseline

**Varying the reduced dimension** We conducted the experiments on text-only model and text+image model to compare the results of varying the reduced dimension. As shown in Figures 8a and 8c, the MSE loss increase when the reduced dimension increase. Figures 8b and 8d shows that the similarity also increased depending on increase of the reduced dimension. We predict that the larger the reduced dimension, the better the mapping of the item and dialog embedding space will be learned due to the higher model complexity. Contrary to our expectations, higher compression of joint embedding space was more important than the model’s complexity for the performance improvement. Moreover, the results implies that the transferring information is more effective when the dimension of implicit joint latent space is small. Our model is efficient in performing multimodal representation learning with a simple network.

**Varying the number of layers** The comparison result of varying the number of layers on the text-only model and text+image model is explained in this section. Figures 9a and 9c illustrate that the MSE loss increased when the depth of TransferNet increases on both models. In addition, MSD similarity also increased with increasing depth in both cases, as shown in Figures 9b and 9d. We tried to improve the performance by increasing the depth, but there was no significant effect. From this result, we conclude that 4 neural blocks are sufficient to learn the mapping between vectors of length 300.

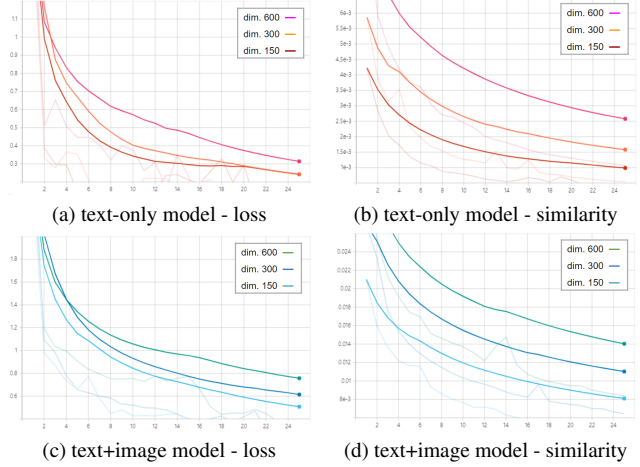


Figure 8. Loss and similarity results of varying the reduced dimension experiment. (a)(b) text-only model (c)(d) text+image model.

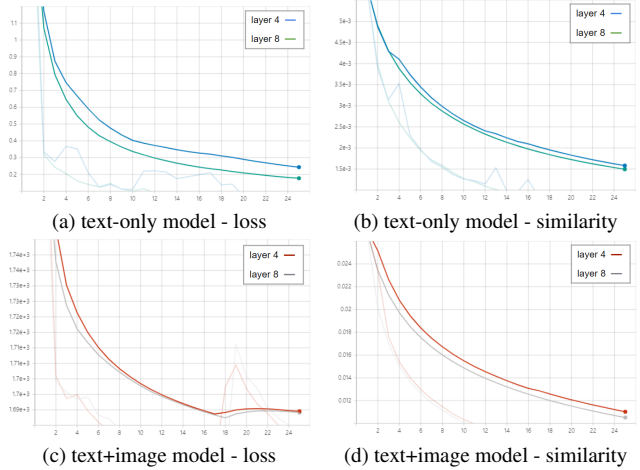


Figure 9. Loss and similarity results of varying the number of layers experiment. (a)(b) text-only model (c)(d) text+image model.

## 6. Discussion

We found that the WKT values are lower than our expectations in analyzing the results while the MSE loss is significantly reduced. The maximum WKT value was 0.1409 when the text+image model is trained with hyperparameters in the default setting. This value implies that our model predicted the ranking with about 30% accuracy. We found several reasons that can explain the failure. First, information about items is highly limited in the dialog. In addition, the evaluation task asks for the prediction of items which do not mention in the conversation. Our current model could not determine whether an item is suitable for a coordi-set because the association of items in the coordi-sets is not learned. In addition, since the embeddings of items that are



in a coordi-set are not located close in the embedding space, the not mentioned items can cause large errors of MSE values unexpectedly.

Furthermore, there are several limitations of our baseline. First, since the used pretrained models did not fine-tune in our domain, there might be a domain gap problem. Domain adaptation should be performed by fine-tuning. In addition, each modality is compressed at different rates, and we used one linear layer for fusion, which is so weak to learn heterogeneity.

Lastly, the prediction of the item ids by the last layer failed. We expected the failure since it is hard to get the correct item id using only text and image information. There should be an additional model that maps joint embedding space to item id space. To deal with this issue, we can use one-hot encoding on embeddings, change the last layer into deep neural networks, or add another loss term such as reconstruction loss from item id to item embedding vector.

## 7. Future Work

To solve upper limitations and develop our recommendation system, we suggest some directions for future work as follows:

**Strengthening fusion network** We used one linear layer in our model to fuse item image vectors and item text vectors. But we recognize that the fusion layer cannot learn the correspondence between item images and descriptions. Using attention mechanisms or transformers can help to learn highly complex relationships, which might improve the performance of predictions.

**Reflect on the degree of harmonizing** To address one of the limitations mentioned in the discussion section, the degree of harmonizing of each item can be judged and used for a better recommendation. The association degree of items can be captured by training a neural network with a dataset that contains many good examples of coordi-sets. Then we can predict the rest more accurately when some of the coordi-set items are decided and this network will help our model to predict the last few items that do not describe in the conversation more accurately.

**Contrastive learning** After different augmentation is applied to the data, the feature representations of two positive pairs are trained to be close while the ones of two negative pairs are trained to be far away from each other in contrastive learning. To apply contrastive learning to our task, we can make positive pairs with similar fashion items. Similar fashion items are items in a coordi-set or items that have similar visual features. The similarity between items can be measured by calculating the distance between items, and this may improve our model's recommendation performance.

**Generative model** We can use Generative Adversarial Networks(GANs) to generate an appropriate coordi-set

from the description or conversation. It can be expected to learn the distribution of items and coordi-sets and propose a set of items by changing a latent vector.

**Continual learning** Currently, our model only relies on the consistent data given in the training process. If we use continual learning to update the model so that it can recommend coordi-sets by reflecting individual user preferences, we expect to get more detailed and accurate results.

## 8. Conclusion

In this paper, we introduce a fast and simple recommendation system for fashion-item which exploit multimodal data. Our model aims to learn a joint embedding space that maps the heterogeneous embedding spaces. We preprocessed FASCODE dataset which has dialog between user and coordinator into three dictionaries for embeddings of dialogues, item images, and item description by using the pretrained KoELECTRA model and the Image2Vec model. Then we designed and rectified our own baseline to predict coordi-sets from dialog and evaluation algorithm to measure the accuracy of our predictions by ranking the candidate coordi-sets. Furthermore, we performed various performance improvement research to figure out the effects of hyperparameters and neural architectures of our Transfer network. The experimental results on four tasks show that our model (1) learns the joint embedding space so that it utilize image information very well, (2) shows strength when some of one modality is lost, (3) is efficient so that a light network is needed. However, it presents some limitations such as low WKT value and weak fusion network. Future research on adding other methods such as contrastive loss or generative model might investigate the association between text and image embedding spaces in fashion domain.

## References

- [1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2
- [2] Euisok Chung, Hyun Woo Kim, Hyo-Jung Oh, and Hwa Jeon Song. Dataset for interactive recommendation system. In *Annual Conference on Human and Language Technology*, pages 481–485. Human and Language Technology, 2020. 2
- [3] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [5] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 30, 2016. 1

- [6] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020. 1
- [7] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013. 2
- [8] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2
- [9] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2
- [10] Jangwon Park. Koelectra: Pretrained electra model for korean. <https://github.com/monologg/KoELECTRA>, 2020. 2
- [11] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012. 1
- [12] Christian Safka. img2vec. 2
- [13] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, page 3, 2012. 2
- [14] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019. 1
- [15] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445, 2019. 1
- [16] J. Weston, S. Bengio, and N. Usunier. Web scale image annotation: Learning to rank with joint word-image embeddings image annotation. *ECML*, 2010. 2
- [17] Wei Yinwei, Wang Xiang, Nie Liqiang, He Xiangnan, and Chua Tat-Seng. Grcn: Graph-refined convolutional network for multimedia recommendation with implicit feedback. *arXiv preprint arXiv:2111.02036*, 2021. 1