10-2021

Multi-modal recommender systems: Hands-on exploration

Quoc Tuan TRUONG
Singapore Management University, qttruong.2017@phdis.smu.edu.sg

Aghiles SALAH
Singapore Management University, asalah@smu.edu.sg

Hady Wirawan LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Multi-Modal Recommender Systems: Hands-On Exploration

QUOC-TUAN TRUONG, Singapore Management University, Singapore
AGHILES SALAH, Rakuten Institute of Technology, France
HADY W. LAUW, Singapore Management University, Singapore

Recommender systems typically learn from user-item preference data such as ratings and clicks. This information is sparse in nature, i.e., observed user-item preferences often represent less than 5% of possible interactions. One promising direction to alleviate data sparsity is to leverage auxiliary information that may encode additional clues on how users consume items. Examples of such data (referred to as modalities) are social networks, item's descriptive text, product images. The objective of this tutorial is to offer a comprehensive review of recent advances to represent, transform and incorporate the different modalities into recommendation models. Moreover, through practical hands-on sessions, we consider cross model/modality comparisons to investigate the importance of different methods and modalities. The hands-on exercises are conducted with Cornac (https://cornac.preferred.ai), a comparative framework for multimodal recommender systems. The materials are made available on https://preferred.ai/recsys21-tutorial/.

ACM Reference Format:
Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. 2021. Multi-Modal Recommender Systems: Hands-On Exploration. In *Fifthteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460231.3473324

1 TOPIC AND OBJECTIVES

1.1 Overview of Preference Models

The foundation to any recommender system is the preference model, i.e., transforming the observations on user-item interactions such as ratings and clicks into predictions whether a user is likely to prefer an item. The predominant approach is to train a model from the preference data. Broadly speaking, there are two main families of models.

Matrix Factorization. The first family is that of matrix factorization [10]. The preference data is represented as a user-by-item matrix, which is then factorized into a set of K-dimensional user and item latent factors. A prediction for a user u and an item i is estimated based on the inner product of u and i's latent factors. This formulation gives rise to variants due to different loss functions. Models based on *explicit feedback* typically seek to minimize the error between the observed and the predicted ratings, as exemplified by PMF [20]. Alternatively, *implicit feedback* models may interpret observations as confidence signal (e.g., WMF [7]) or as relative comparisons (e.g., BPR [24]).

Neural Networks. The second family is that of neural networks, which incorporates non-linearity through activation functions and multiple compositional layers. There are two main treatments of preference data. For one, target ratings are used as supervision to learn the weights of a multi-layer perceptron (MLP), as exemplified by NCF [6]. For another, we learn a lower-dimensional representation of observed interactions with the help of auto-encoders, which upon decoding produce predictions on other interactions yet to be observed, as exemplified by VAE-CF [15] and BiVAE [31].

Manuscript submitted to ACM

1.2 Multi-Modality and Cornac

For the hands-on sessions we rely on Cornac, which is a Python framework tailored for multimodal recommender systems [28]. In addition to essential features for building, accessing, benchmarking and evaluating recommender models, the hallmark of Cornac is its native *multimodal* support. This is reflected by Cornac's rich collection of recommender models integrating various auxiliary data types, namely *text*, *graph*, and *image*, as well as standardized pipelines for handling these modalities. Besides making it convenient to work with auxiliary data, one key aspect of this standardization is broadening the use case of existing models by enabling *cross-modality* utilization, which we explore in the last of part of this tutorial. For a more conceptual introduction and a general overview of Cornac's main modules please refer to the following paper [28]. For a more practical experience, Cornac's GitHub repository (https://github.com/PreferredAI/cornac) includes examples and tutorials illustrating the framework possibilities.

1.3 Text Modality

One way to categorize multi-modal recommender systems based on text modality is by how text is being represented. Table 1 illustrates a selection of representative text-based multi-modal algorithms to be included in this tutorial, as organized by which text representation model is combined with which family of preference models.

Table 1. A selection of text-based multi-modal recommender models to be discussed in the tutorial

Text Representation	Preference Model		
	Explicit MF	Implicit MF	Neural Networks
Term Vector	CCF [16]		
Matrix Factorization	CMF [29]		
Topic Model	HFT [19]	CTR [33], CTRank [37]	
Auto-Encoder	AutoSVD++ [40]	CDL [34], CVAE [14], CDR [38]	
CNN	ConvMF [9], DeepCoNN [42]		
RNN			MRG [30], NRT [13]

Term Vector The simplest way to represent text is as a term vector, where each element indicates the importance of a word. CCF [16] integrates content-based similarity into a matrix factorization model for preference data.

Matrix Factorization Instead of term vectors with dimensionality of vocabulary size, we can model a lower-dimensional representation via matrix factorization on document-term matrix. CMF [29] joins the two matrix factorizations, for text and preference data respectively, by tying the latent vector that represents an item/document.

Topic Model For semantic interpretability, a popular way to model a corpus is a topic model such as LDA [1]. Each document is associated with a probability distribution over topics, and each topic with a probability over words. The integration of LDA with PMF yields HFT [19], with WMF yields CTR [33], and with BPR yields CTRank [37].

Auto-Encoder Neural topic models are based on auto-encoders. AutoSVD++ relies on contractive auto-encoders, CDL [34] and CDR [38] on stacked denoising auto-encoders, and CVAE [14] on variational auto-encoders.

Convolutional Neural Network Topic models typically do not retain the sequence of words in text. One way to do so is through Convolutional Neural Network (CNN). For instance, ConvMF [9] uses CNN to derive item representations from text reviews. DeepCoNN [42] uses CNN on both items' as well as users' reviews.

Recurrent Neural Network In some scenarios, we may seek to generate a sequential text via Recurrent Neural Networks (RNN) or one of its variants. For instance, NRT [13] and MRG [30] integrate Long Short-Term Memory (LSTM) for text generation with a multi-layer perceptron for rating prediction.

1.4 Image Modality

On e-commerce applications, products are often illustrated with descriptive images. Much as with text, if not even more so, images are high-dimensional, dense and thus difficult to incorporate rawly into recommender systems. There are two prominent approaches in representing images, either by extracting low-dimensional embedding from pre-trained neural networks on large vision datasets (e.g., ILSVRC), or learning a convolutional neural network module to extract visual features directly from pixel level. The perceptual information then can be infused with preference signals under the matrix factorization framework. Table 2 organizes some representative image-based multi-modal algorithms by how the image is being represented and which family of models are being used to learn the preferences.

Table 2. A selection of image-based multi-modal recommender models to be discussed in the tutorial

Image Representation	Preference Model	
	Explicit MF	Implicit MF
Pre-trained Embedding	VMF [23], VPOI [35]	VBPR [5], ACF [3], NPR [22]
Convolutional Neural Networks		DVBPR [8], CDL [12], CKE [39], JRL [41]

Pre-trained Embedding Dense embedding vectors of images can be obtained from pre-trained deep neural networks. They are further reduced to lower dimensions, with learned projection, to fit into the bilinear matrix factorization framework. For instance, VMF [23] leverages this approach with explicit MF for learning preference from rating values, where VPOI [35] applies the same technique to POI check-in data. Similarly to VBPR [5], ACF [3] employs BPR for preference ranking and successfully applies to video data, in addition to images. NPR [22] further augments visual preference with spatial and topical signals linked to image metadata.

Convolutional Neural Networks In place pre-trained embeddings, this family of models learn a CNN module to extract visual features from pixel level. DVBPR [8] is a direct extension of VBPR, while CDL [12] employs MLP to learn a more holistic user representation by leveraging other user metadata (e.g., tags). CKE [39] and JRL [41] treat visual data as one view in a multi-view learning approach for user/item representation.

1.5 Graph Modality

Many auxiliary information arising in recommender system applications are graph in nature. On the user-side, social networks are typical examples. On the item side, these are information representing item-relatedness such as product networks and knowledge graphs. Compared to the text and image modalities, the graph modality has the flexibility of encoding signals beyond feature similarities (e.g., complementary products), which can often point towards interesting items that do not match users' typical preferences. Here we focus on important real-world graph auxiliary data and the major families of models, depicted in Table 3, for leveraging this information in recommender systems.

Table 3. A selection of graph-based multi-modal recommender models we consider in the tutorial

Family	Graph Type	
	Social Network	Item Graph
Feature-based	Sorec [17], CVAE, JVAE [11], GraphRec [4]	MCF [23], PCRL [26]
Regularization-based	SoReg [18], Soc-movMF [27]	
Architecture-based	SPF [2], SocialRBM [21]	C2PF [25], KGAT [36]

Feature-based Similarly to text and image, one can derive low-dimensional representations of users/items from graph data that can be leveraged by the preference model. Many models in the literature fall into this category such as Sorec [17], MCF [23], PCRL [26], Conditional/Joint VAEs [11], GraphRec [4]. Differences are in model assumptions and in how the graph is integrated to the preference model.

Regularization-based These methods regularize the latent space of the preference model using the graph modality, i.e., typically in such a way as to encourage connected users/items to have similar latent representations [18, 27].

Architecture-based In this family, the graph information is reflected in the preference model's architecture. Representative examples that we discuss in this tutorial are SPF [2], SocialRBM [21], C2PF [25], and KGAT [36].

1.6 Cross-Modality Utilization

As we go through the three distinct types of modalities and the respective algorithms to deal with them, a keen observer would see that in some cases there exist sufficient commonalities in the way that modalities are represented such that an algorithm that was originally designed for one modality may turn out to also be capable of accommodating a different modality. Such cross-modality utilization may bring a potential benefit in broadening the set of applicable algorithms for a given modality [32]. With datasets that support multiple modalities, we explore this further in several directions. For one, we see how different modalities may respectively contribute to recommendations. For another, we pay attention to whether a model designed for one modality could potentially perform better with another modality instead. This exploration would inform whether we should perceive and develop multimodal recommender systems in separate modality streams, or we should approach multimodality in a more holistic and inter-operable manner.

2 SCHEDULE

The tutorial length is 180 minutes. Of these, 90 minutes are to cover the classic and recent literature, and 90 minutes are for hands-on coding exercises with Jupyter notebook. Concretely, the tutorial has the following schedule:

(1) Brief overview of recommender systems (20 minutes).
(2) Introduction to multimodal recommender systems (20 minutes).
(3) Hands-on: Starting with the Cornac framework (10 minutes).
(4) Exploration into each modality (90 minutes).
(5) Cross-modal utilization (30 minutes).
(6) Future directions (10 minutes).

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3 (2003), 993–1022.
[2] Allison JB Chaney, David M Blei, and Tina Eliassi-Rad. 2015. A probabilistic model for using social networks in personalized item recommendation. In *RecSys*. 43–50.
[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*. 335–344.
[4] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *WWW*. 417–426.
[5] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*.

[6] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[7] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM*. 263–272.

[8] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *ICDM*. 207–216.

[9] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional matrix factorization for document context-aware recommendation. In *RecSys*. 233–240.

[10] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[11] Wonsung Lee, Kyungwoo Song, and Il-Chul Moon. 2017. Augmented variational autoencoders for collaborative filtering with auxiliary information. In *CIKM*. 1139–1148.

[12] Chenyi Lei, Dong Liu, Weiping Li, Zheng-Jun Zha, and Houqiang Li. 2016. Comparative deep learning of hybrid representations for image recommendations. In *CVPR*. 2545–2553.

[13] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*.

[14] Xiaopeng Li and James She. 2017. Collaborative variational autoencoder for recommender systems. In *SIGKDD*.

[15] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *WWW*.

[16] Zhongqi Lu, Zhicheng Dou, Jianxun Lian, Xing Xie, and Qiang Yang. 2015. Content-based collaborative filtering for news topic recommendation. In *AAAI*.

[17] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. 2008. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*. 931–940.

[18] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *WSDM*. 287–296.

[19] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*. 165–172.

[20] Andriy Mnih and Russ R Salakhutdinov. 2008. Probabilistic matrix factorization. In *NIPS*. 1257–1264.

[21] Trong T Nguyen and Hady W Lauw. 2016. Representation learning for homophilic preferences. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 317–324.

[22] Wei Niu, James Caverlee, and Haokai Lu. 2018. Neural personalized ranking for image recommendation. In *WSDM*. 423–431.

[23] Chanyoung Park, Donghyun Kim, Jinoh Oh, and Hwanjo Yu. 2017. Do" Also-Viewed" Products Help User Rating Prediction?. In *WWW*. 1113–1122.

[24] Rendle S., Freudenthaler C., Gantner Z., and Schmidt-Thieme L. 2012. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*.

[25] Aghiles Salah and Hady W Lauw. 2018. A bayesian latent variable model of user preferences with item context. IJCAI.

[26] Aghiles Salah and Hady W Lauw. 2018. Probabilistic collaborative representation learning for personalized item recommendation. In *UAI*.

[27] Aghiles Salah and Mohamed Nadif. 2017. Social regularized von Mises–Fisher mixture model for item recommendation. *Data Mining and Knowledge Discovery* 31, 5 (2017), 1218–1241.

[28] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *JMLR* 21 (2020), 95–1.

[29] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *SIGKDD*. 650–658.

[30] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal review generation for recommender systems. In *WWW*. 1864–1874.

[31] Quoc-Tuan Truong, Aghiles Salah, and Hady W Lauw. 2021. Bilateral Variational Autoencoder for Collaborative Filtering. In *WSDM*. 292–300.

[32] Quoc-Tuan Truong, Aghiles Salah, Thanh-Binh Tran, Jingyao Guo, and Hady W. Lauw. 2021. Exploring Cross-Modality Utilization in Recommender Systems. *IEEE Internet Computing* (2021).

[33] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *SIGKDD*. 1100–1108.

[34] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *SIGKDD*.

[35] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. 2017. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *WWW*. 391–400.

[36] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *SIGKDD*. 950–958.

[37] Weilong Yao, Jing He, Hua Wang, Yanchun Zhang, and Jie Cao. 2015. Collaborative topic ranking: Leveraging item meta-data for sparsity reduction. In *AAAI*.

[38] Haochao Ying, Liang Chen, Yuwen Xiong, and Jian Wu. 2016. Collaborative deep ranking: A hybrid pair-wise recommendation algorithm with implicit feedback. In *PAKDD*.

[39] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *SIGKDD*. 353–362.

[40] Shuai Zhang, Lina Yao, and Xiwei Xu. 2017. Autosvd++ an efficient hybrid collaborative filtering model via contractive auto-encoders. In *SIGIR*.

[41] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *CIKM*. 1449–1458.

[42] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*.