The ACM Conference Series on Recommender Systems

L14: Multimodal Deep Learning for Recommendation

ACM RecSys Summer School 2024

Daniele Malitesta, PhD

Université Paris-Saclay, CentraleSupélec, Inria

October 09, 2024

CentraleSupélec

université PARIS-SACLAY

de' remi facemmo ali' 1990 POLITECNICO DI BARI

Who's in front of you

- Postdoc researcher, Université Paris-Saclay, CentraleSupélec, Inria (Gif-sur-Yvette, France)
- PhD in Computer Science Engineering at Politecnico di Bari (Bari, Italy)
- Born and raised in Bari!
- PhD on GNNs-based approaches for recommendation leveraging multimodal information
- Currently working on graph machine learning, fairness, generative models
- Email: d.malitesta@gmail.com

Lecture schedule

- **Part 0:** I wish I had known it in advance! [5 min.]
  - *0.1: (Some) useful resources to get started*
  - *0.2: Field-specific conferences/journals*
- **Part 1:** One multimodal schema to rule them all [50 min.]
  - *1.1: Personalized (multimedia) recommendation*
  - *1.2: Why do we need a schema anyway?*
  - *1.3: [Theory + Practice] A unified multimodal schema*
- **Part 2:** Open challenges and how we solved them [15 min.]
  - *2.1: Missing modalities in the input data*
  - *2.2: Multimodality on user-item interactions*
  - *2.3: Pre-trained feature extractors*
  - *2.4: Fine-grained multimodal features*
  - *2.5: An extensive and fair evaluation*
- **Part 3:** Q & A time [$+\infty$[1]]

[1] just kidding :-D

Lecture resources

All useful resources regarding this lecture are made available here:
`https://github.com/danielemalitesta/Multimodal-DL-4-RecSys`.
If you like it and find it useful, consider putting a star :-)

Scan me!

Part 0: I wish I had known it in advance!

(Some) useful resources to get started

Multimodal deep learning

- Multimodal Machine Learning: A Survey and Taxonomy (Baltrusaitis et al. [4])
- Multimodal Deep Learning (Ngiam et al. [15])
- LTI-11777: Multimodal Machine Learning (course website)
- PyTorch (link)
- TensorFlow (link)
- Hugging Face (link)
- TorchMultimodal (link)

Multimodal-aware recommendation

- Formalizing Multimedia Recommendation through Multimodal Deep Learning (Malitesta et al. [12])
- Ducho Meets Elliot: Large-scale Benchmarks for Multimodal Recommendation
- Multimodal Recommender Systems: A Survey (Liu et al. [11])
- A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions (Zhou et al. [26])
- Multi-Modal Recommender Systems: Hands-On Exploration (Truong et al. [20])
- Ducho + Elliot (link)
- Cornac (link)
- MMRec (link)

Legend: papers tutorials frameworks courses

Field-specific conferences/journals

Part 1: One multimodal schema to rule them all

Personalized recommendation

"Recommender Systems are software tools and techniques providing suggestions for items to be of use of a user. The suggestions provided are aimed at supporting their users in various decision-making processes, such as items to buy, what music to listen , or what news to read [18]."

Recommender systems built on MF

Most recommendation approaches builds on factorization models, such as matrix factorization (MF) [9]. They map users and items to embeddings and reconstruct the interaction matrix through their dot-product.

Inexplicable nature of implicit feedback

In implicit feedback settings, it is not easy to profile users since it is not always clear whether an interaction may represent actual preference. This is even more evident when users interact with items with multi-faceted characteristics (e.g., multimedia content such as photos, videos, songs, texts, etc.).

Multimedia recommendation

In domains such as fashion [7], song [16], micro-videos [22], and food [14] recommendation, multimedia recommender systems augment the representation of items in the catalogue through their multimodal features extracted from images, texts, or audio tracks describing them.

Multimodal deep learning

Multimodal deep learning have addressed applicative domains such as medical imaging [19], autonomous driving [5], speech/emotion recognition [17], multimedia retrieval [10], and, only recently, multimodal large language modelling [23]. Some works have tried to outline, categorize, and formalize the core concepts behind multimodality in deep learning [3] through a pipeline: *representation*, *translation*, *alignment*, *fusion*, and *co-learning*.

Bridging multimodal DL and multimedia recommendation

However, and differently from the other outlined domains and scenarios, recommendation lacks a *shared* theoretical and applicative formalization to align the multimedia recommendation problem with the same formal pipeline proposed in multimodal deep learning.

Classical recommendation

We denote with $u \in \mathcal{U}$, $i \in \mathcal{I}$, and $r \in \mathcal{R}$ a user, an item, and a user-item interaction, respectively. To ease the notation, we say $x \in \mathcal{X}$ is a general input to the system, with $\mathcal{X} = \mathcal{U} \cup \mathcal{I} \cup \mathcal{R}$. Given a set of input data $\mathcal{X}$, and defined $\rho(\cdot)$ as the preference score prediction function, a recommender system aims to build a top@k list of items maximizing the following posterior probability ($prob$):

$$\hat{\Theta}_\rho = \arg\max_{\mathbf{\Theta}_\rho} prob(\mathbf{\Theta}_\rho \mid \mathcal{X}), \tag{1}$$

where $\mathbf{\Theta}_\rho = [\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}]$ is the vector collecting all weights for the inference function $\rho(\cdot)$, $\mathcal{W}_\rho = \{\theta_\rho^{(0)}, \theta_\rho^{(1)}, \ldots, \theta_\rho^{(|\mathcal{W}_\rho|-1)}\}$ is the set of such weights, and $|\mathcal{W}_\rho|$ its cardinality. Thus, in the case of classical recommender system leveraging latent factor models, the inference function may be indicated as:

$$\hat{y} = \rho(\theta_\rho^{\mathcal{U}}, \theta_\rho^{\mathcal{I}}, \mathcal{X}), \tag{2}$$

where $\hat{y}$ is the predicted value through the inference function $\rho(\cdot)$, such as a rating score for any user-item pair.

The multimodal schema

Multimodal pipeline as introduced in [12].

Multimodal data input

Formally, we define $m \in \mathcal{M}$ as an admissible modality for the system (i.e., $\mathcal{M} = \{\mathbf{v}, \mathbf{t}, \mathbf{a}\}$). Let $x \in \mathcal{X}$ be an input to the recommender system, whose set of available modalities is indicated as $\mathcal{M}_x \subseteq \mathcal{M}$. We represent the *content* data of input x in modality m as $c_x^{(m)}$, with $m \in \mathcal{M}_x$, and the vector of content data for input x in all modalities as c_x.

Multimodal feature processing

Feature extraction. Our schema introduces a *Feature Extractor* (FE) component to extract features. Let $c_x^{(m)}$ be the content data for input x in modality $m \in \mathcal{M}_x$. Then, let $\varphi_m(\cdot)$ be the feature extractor function for the modality m. We define the feature extraction process in the modality m as:

$$\overline{c}_x^{(m)} = \varphi_m(c_x^{(m)}) \quad \forall m \in \mathcal{M}_x, \tag{3}$$

where $\overline{c}_x^{(m)}$ is the extracted feature for input x in modality m. We use the notation $\overline{\mathrm{c}}_x = [\overline{c}_x^{(0)}, \overline{c}_x^{(1)}, \ldots, \overline{c}_x^{(|\mathcal{M}_x|-1)}]$ to refer to the vector of extracted features for input x in all modalities.

A visual-aware feature extractor as investigated in [6].

Multimodal feature processing (contd.)

Multimodal representation. The next step is to design a *Representation* strategy to handle the relationships among modalities and eventually inject such data into the recommender system. The literature follows two main approaches: *Joint* and *Coordinate*. Whatever the chosen strategy, the final multimodal representation is indicated as $\tilde{c}_x$.

Joint and coordinate representation as presented in [12].

Multimodal feature fusion

As an optional third step, when *Coordinate* representation is used, our multimodal schema allows an additional *Fusion* step to combine all produced multimodal representations.

Early fusion. Let $\tilde{\mathrm{c}}_x$ be the vector of coordinate multimodal representations for input x in all modalities. Then, let $\gamma_e(\cdot)$ be the function for *Early* fusion. We generate the multimodal representation for input x as:

$$\tilde{c}_x = \gamma_e(\tilde{\mathrm{c}}_x). \tag{4}$$

We obtain the predicted output $\hat{y}$ for input x as:

$$\hat{y} = \rho(\tilde{c}_x). \tag{5}$$

Multimodal feature fusion (contd.)

Late fusion. Let $\tilde{c}_x^{(m)}$ be the coordinate multimodal representation for input x in modality $m \in \mathcal{M}_x$. We first predict the different output values for each modality as:

$$\hat{y}^{(m)} = \rho(\tilde{c}_x^{(m)}) \quad \forall m \in \mathcal{M}_x. \tag{6}$$

Let $\widehat{\mathrm{y}}$ be the vector of multimodal predicted outputs in all modalities. If we denote $\gamma_l(\cdot)$ as the function for *Late* fusion, we finally aggregate (fuse) all modality-aware predictions:

$$\hat{y} = \gamma_l(\widehat{\mathrm{y}}). \tag{7}$$

The final task

Let $\mathcal{W}_\varphi$, $\mathcal{W}_\mu$, and $\mathcal{W}_\gamma$ be the sets of the additional model trainable weights from (i) feature extraction, (ii) multimodal representation, and (iii) multimodal fusion, respectively. Note that they could be empty, as the correspondent functions may be non-trainable. Then, given the set of multimodal input data $\mathcal{X}$:

$$\hat{\Theta} = \arg\max_{\mathbf{\Theta}} prob(\mathbf{\Theta}|\mathcal{X}), \tag{8}$$

where $\mathbf{\Theta} = [\mathbf{\Theta}_\rho, \mathbf{\Theta}_\varphi, \mathbf{\Theta}_\mu, \mathbf{\Theta}_\gamma]$. Thus, in the case of multimodal recommender systems leveraging latent factor models, the inference function may be indicated:

$$\hat{y} = \rho(\theta^{\mathcal{U}}_\rho, \theta^{\mathcal{I}}_\rho, \theta^{\mathcal{U}}_{\mu_m}, \theta^{\mathcal{I}}_{\mu_m}, \mathcal{X}), \quad \forall m \in \mathcal{M}. \tag{9}$$

Hands-on session

Let us take a look at VBPR [8], the most standard multimedia recommendation model, in its multimodal version (as introduced in this work [24]). The architecture takes the BPR-MF model and adds the multimedia/multimodal part regarding users and (most importantly) items.

$$\hat{y} = \left(\theta_{\rho}^{\mathcal{U}}\right)^{\top} \theta_{\rho}^{\mathcal{I}} + \left(\theta_{\mu_m}^{\mathcal{U}}\right)^{\top} \mu(\overline{c}_x^{(m_1)}, \overline{c}_x^{(m_2)}, \ldots, \overline{c}_x^{(m_{|\mathcal{M}|})}), \quad \forall m \in \mathcal{M}, \tag{10}$$

where we have:

- $\theta_{\rho}^{\mathcal{U}}$, $\theta_{\rho}^{\mathcal{I}}$ $\rightarrow$ collaborative user and item embeddings
- $\theta_{\mu_m}^{\mathcal{U}}$ $\rightarrow$ multimodal representation user embeddings
- $\mu(\overline{c}_x^{(m_1)}, \overline{c}_x^{(m_2)}, \ldots, \overline{c}_x^{(m_{|\mathcal{M}|})})$ $\rightarrow$ *Joint* multimodal representation of items extracted features (concatenation + projection)[2]

[2] Formally, *Coordinate* representation (projection) + *Early* fusion (concatenation).

Hands on session (contd.)

Let's go to Google Colab! You can find it in the GitHub repository shared at the beginning of this lecture!

Scan me!

Part 2: Open challenges and how we solved them

Missing modalities in input data

The need to provide descriptive content for every input modality may come at the expense of some missing modalities (e.g., product photos or descriptions could be unavailable). The issue is very poorly debated in the literature. However, in our recent work, we provided a graph-based solution to address it, acting as a pre-processing procedure [13].

Multimodality on user-item interactions

Existing review-based approaches work by integrating reviews as the *textual modality* to represent *items* [25, 21]. However, we believe that a more logical and effective way to integrate reviews would be to view them as a medium to represent *user* preference *over items*, thereby providing additional and complementary preference scores in addition to numeric ratings or implicit feedback that are typically used to compute recommendations [1].

Pre-trained feature extractors

Two main issues: (i) feature extractors are used as pre-trained blocks, but this does not mean the pre-trained knowledge will be helpful for recommendation; (ii) the literature enumerates always the same configuration of multimodal extractors. In our recent benchmarking work, we analyzed these two aspects [2].

Models	**Extractors**	**Toys & Games**		
		Recall	nDCG	HR
VBPR	RNet50 + SBert	10.83	5.70	16.54
	MMF + SBert	10.80	5.72	16.43
	Clip	<u>10.98</u>	<u>5.83</u>	<u>16.80</u>
	Align	**11.06**	**5.85**	**16.86**
	AltClip	10.92	<u>5.83</u>	16.60

Fine-grained multimodal features

In contexts where modalities bring a great source of heterogeneous information, a finer-grained feature processing could help better unveil hidden facets. For instance, when it comes to the recommendation of fashion items (e.g., dresses, shoes, jewelry), user attention may be captivated by specific item visual characteristics, such as colors, shapes, and particular patterns and motifs.

Content-style item representation as proposed in [7].

An extensive and fair evaluation

We have just released a large-scale benchmarking analysis on multimodal recommendation [2]. Currently, we are hosting a leaderboard on **Papers with Code**. Feel free to contribute!

Scan me!

Part 3: Q & A time

Many people to thank!

None of the things I said and wrote in these slides would have been possible without the great contribution of former and current collaborators/supervisors.

I wish to thank them all!

Supervisors (former and current): Tommaso Di Noia (Poliba), Fragkiskos Malliaros (CentraleSupélec).

Collaborators (alphabetical order): Walter Anelli (Poliba), Matteo Attimonelli (Poliba), Giandomenico Cornacchia (IBM Europe) Danilo Danese (Poliba), Angela Di Fazio (Poliba), Antonio Ferrara (Poliba), Giuseppe Gassi (Poliba), Alberto Mancino (Poliba), Felice Merra (AWS GenAI), Claudio Pomo (Poliba), Emanuele Rossi (VantAI).

Anonymous survey on the lecture

If you have attended the lecture, please leave a comment in this **anonymous form**. It will take a few minutes, but it can be very important to me to improve the quality of the lecture! Thank you for your time and contribution!

Scan me!

References I

[1] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, Antonio Ferrara, Daniele Malitesta, and Claudio Pomo. Reshaping graph recommendation with edge graph collaborative filtering and customer reviews. In *DL4SR@CIKM*, volume 3317 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[2] Matteo Attimonelli, Danilo Danese, Angela Di Fazio, Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. Ducho meets elliot: Large-scale benchmarks for multimodal recommendation. *CoRR*, abs/2409.15857, 2024.

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Challenges and applications in multimodal machine learning. In *The Handbook of Multimodal-Multisensor Interfaces, Volume 2 (2)*, pages 17–48. Association for Computing Machinery, 2018.

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

References II

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628. Computer Vision Foundation / IEEE, 2020.

[6] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. A study on the relative importance of convolutional neural networks in visually-aware recommender systems. In *CVPR Workshops*, pages 3961–3967. Computer Vision Foundation / IEEE, 2021.

[7] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. Leveraging content-style item representation for visual recommendation. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 84–92. Springer, 2022.

[8] Ruining He and Julian J. McAuley. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150. AAAI Press, 2016.

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

References III

[10] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. Hybrid fusion with intra- and cross-modality attention for image-recipe retrieval. In *SIGIR*, pages 244–254. ACM, 2021.

[11] Qidong Liu, Jiaxi Hu, Yutian Xiao, Jingtong Gao, and Xiangyu Zhao. Multimodal recommender systems: A survey. *CoRR*, abs/2302.03883, 2023.

[12] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. Formalizing multimedia recommendation through multimodal deep learning. *ACM Trans. Recomm. Syst.*, apr 2024. doi: 10.1145/3662738. URL `https://doi.org/10.1145/3662738`. Just Accepted.

[13] Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D. Malliaros. Do we really need to drop items with missing modalities in multimodal recommendation? *CoRR*, abs/2408.11767, 2024.

[14] Weiqing Min, Shuqiang Jiang, and Ramesh C. Jain. Food recommendation: Framework, existing solutions, and challenges. *IEEE Trans. Multim.*, 22(10): 2659–2671, 2020.

References IV

[15] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696. Omnipress, 2011.

[16] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. A deep multimodal approach for cold-start music recommendation. In *DLRS@RecSys*, pages 32–37. ACM, 2017.

[17] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *ACL (1)*, pages 4491–4503. Association for Computational Linguistics, 2022.

[18] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer, 2015.

[19] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. MATR: multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.*, 31:5134–5149, 2022.

References V

[20] Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. Multi-modal recommender systems: Hands-on exploration. In *RecSys*, pages 834–837. ACM, 2021.

[21] Xi Wang, Iadh Ounis, and Craig Macdonald. Leveraging review properties for effective recommendation. In *WWW*, pages 2209–2219. ACM / IW3C2, 2021.

[22] 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 *ACM Multimedia*, pages 1437–1445. ACM, 2019.

[23] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023.

[24] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *ACM Multimedia*, pages 3872–3880. ACM, 2021.

References VI

[25] Lei Zheng, Vahid Noroozi, and Philip S. Yu. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*, pages 425–434. ACM, 2017.

[26] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *CoRR*, abs/2302.04473, 2023.