The 2nd Edition of EvalRS: a Rounded Evaluation of Recommender Systems
Long Beach, CA, USA 07-08-2023
KDD 2023 - Workshops

Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems

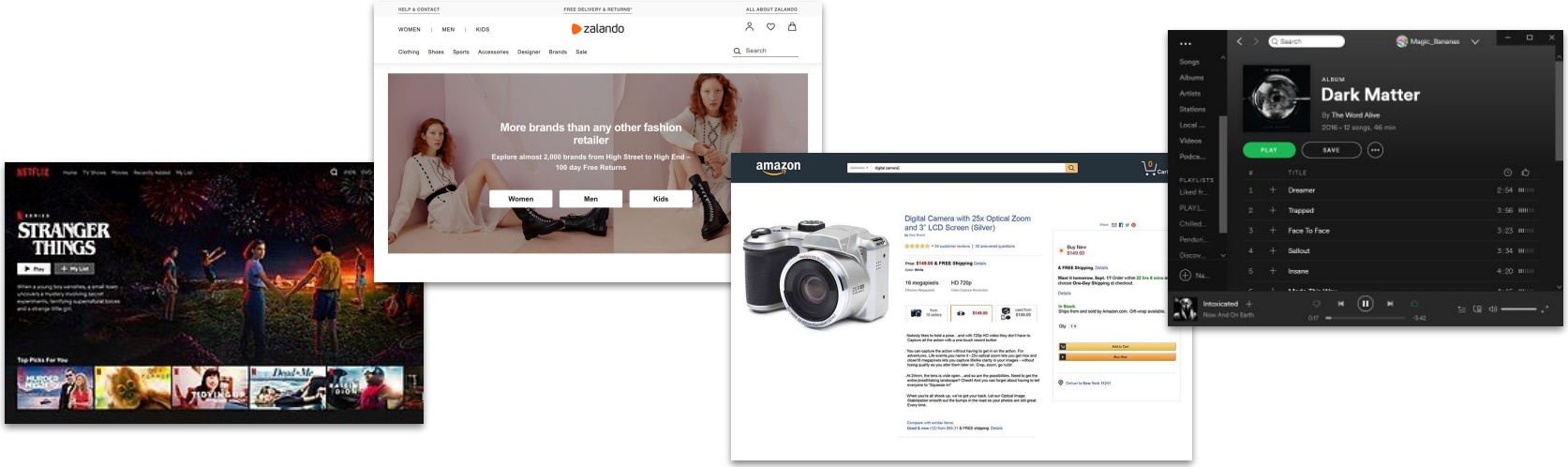
Daniele Malitesta, Giandomenico Cornacchia, *Claudio Pomo*, Tommaso Di Noia

Politecnico di Bari
Bari, Italy
email: firstname.lastname@poliba.it

Introduction and Motivation

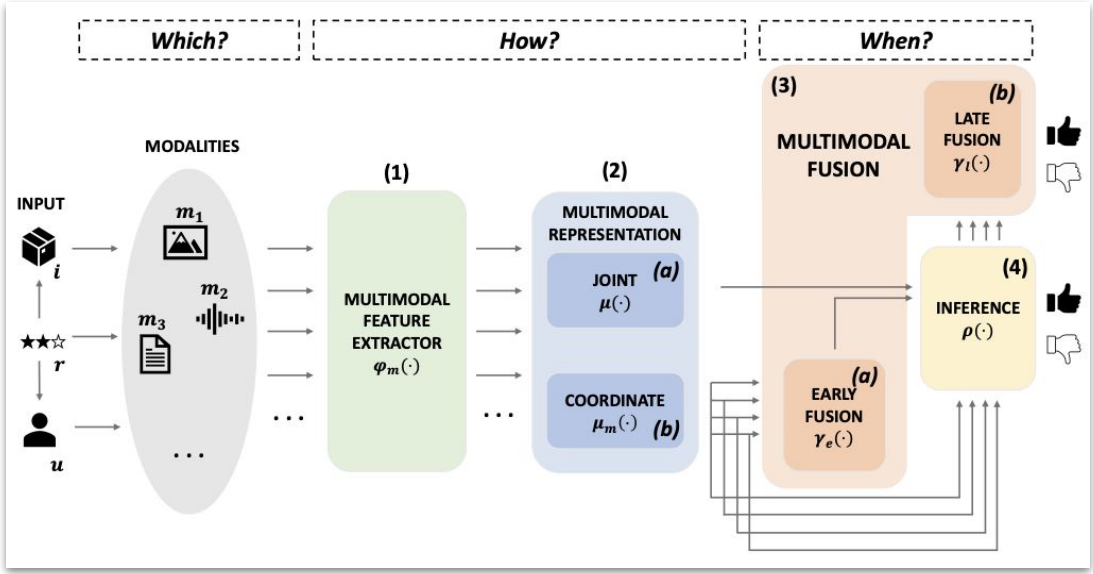
Multimodal-aware recommendation

In domains such as **fashion**, **music**, **food**, and **video** recommendation, recommender systems (RSs) may leverage items' **multimodal** side information (e.g., product images and descriptions, or audio tracks) to tackle (i) data **sparsity**, (ii) **cold-start** scenario, (iii) the **inexplicable** nature of **implicit** users' feedback. Such a family of recommender systems is known as **multimodal-aware** recommender systems (MRSs).



Multimodal-aware recommendation (cont'd)

The typical **multimodal** recommendation **pipeline** consists of four steps: (i) **high-level** features are **extracted** via **pre-trained** deep neural networks, then (ii) multimodal **representations** of users and/or items are **learned**, to optionally (iii) **fuse** them and (iv) **estimate** a user-item interaction **score**.



Politecnico di Bari

SISINF LAB

An evaluation gap in the literature

Despite the **success** of MRSs, performance **concerns** still raise: (i) as most of such approaches propose **slight variations** on a **common** theme (i.e., **matrix factorization** with multimodal content), it is **not** always **clear** which **strategy** is providing the most **significant contribution**, (ii) existing MRSs are **trained** and **evaluated** under **different** implementations and settings.

Our contributions

- Provide a **unified framework** to benchmark **five** state-of-the-art multimodal-aware recommender systems (i.e., VBPR, MMGCN, MGAT, GRCN, LATTICE).
- Run **extensive hyper-parameter explorations** to fine-tune all tested models **under the same settings** for a **fair comparison**.
- To the best of our knowledge, this is **the first attempt** to evaluate MRSs on **measures** accounting for **accuracy**, **novelty**, and **diversity**.

Proposed analysis

Experimental settings and reproducibilty

Models	Venue	Baseline in
VBPR [9]	AAAI 2016	[20, 43, 32, 11, 13, 44]
MMGCN [1]	MM 2019	[39, 45, 46, 21, 47, 48]
MGAT [11]	IPM 2020	[49, 50]
GRCN [12]	MM 2020	[51, 13, 44, 47, 48, 52]
LATTICE [13]	MM 2021	[47, 48, 53, 14, 54]

Datasets	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{R}\rvert$	Sparsity (%)
Office	4,905	2,420	53,258	99.5513
Toys	19,412	11,924	167,597	99.9276
Clothing	39,387	23,033	278,677	99.9693

- 5-core on users and items
- 80%/20% training and test splitting
- 50% of the test used as validation on Recall@20
- Epochs are 200 for all models

Codes, datasets, and configs

Politecnico di Bari

SISINF LAB

Evaluation metrics

Expected Free Discovery [*]

$$\mathrm{EFD@}k = C \sum_{i_k \in R} \mathrm{disc}(k) P\left(\mathrm{Rel}_u\ @k \mid i_k, u\right) \cdot \cdot \left(-\log_2 p(i \mid \text{ seen}, \theta)\right)$$

Item Coverage

$$\mathrm{iCov@}k = \frac{|\bigcup_u \hat{\mathcal{I}}_u @k|}{|\mathcal{I}_{train}|}$$

Gini Index [**]

$$\mathrm{Gini@}k = 1 - \left(\frac{\sum_{i=1}^{|\mathcal{I}|} (2i - |\mathcal{I}| - 1) P|_{@k}(i)}{|\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} P|_{@k}(i)} \right)$$

References

[*] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: RecSys, ACM, 2011, pp. 109–116.

[**] W. Sun, S. Khenissi, O. Nasraoui, P. Shafto, Debiasing the human-recommender system feedback loop in collaborative filtering, in: WWW (Companion Volume), ACM, 2019, pp. 645–651.

Results and discussion

Accuracy performance (RQ1)

Datasets	Models	$k = 10$			$k = 20$			$k = 50$		
		Recall	nDCG	Prec	Recall	nDCG	Prec	Recall	nDCG	Prec
Office	VBPR	0.0652	0.0419	0.0164	0.1025	0.0533	0.0133	0.1774	0.0721	0.0095
	MMGCN	0.0455	0.0300	0.0124	0.0798	0.0405	0.0109	0.1575	0.0598	0.0084
	MGAT	0.0427	0.0277	0.0119	0.0745	0.0377	0.0102	0.1450	0.0552	0.0079
	GRCN	0.0393	0.0253	0.0118	0.0667	0.0339	0.0099	0.1250	0.0488	0.0075
	LATTICE	**0.0664**	**0.0449**	**0.0173**	**0.1029**	**0.0566**	**0.0137**	**0.1780**	**0.0751**	**0.0096**
Toys	VBPR	0.0710	0.0458	0.0131	0.1006	0.0545	0.0096	0.1523	0.0667	0.0061
	MMGCN	0.0256	0.0150	0.0052	0.0426	0.0200	0.0044	0.0785	0.0285	0.0033
	MGAT	0.0375	0.0230	0.0072	0.0595	0.0294	0.0059	0.1039	0.0398	0.0043
	GRCN	0.0554	0.0354	0.0108	0.0831	0.0436	0.0083	0.1355	0.0559	0.0056
	LATTICE	**0.0805**	**0.0512**	**0.0148**	**0.1165**	**0.0617**	**0.0110**	**0.1771**	**0.0759**	**0.0069**
Clothing	VBPR	0.0339	0.0181	0.0034	0.0529	0.0229	0.0027	0.0847	0.0292	0.0017
	MMGCN	0.0227	0.0119	0.0023	0.0348	0.0150	0.0018	0.0609	0.0201	0.0012
	MGAT	0.0244	0.0127	0.0025	0.0384	0.0162	0.0019	0.0699	0.0225	0.0014
	GRCN	0.0319	0.0164	0.0032	0.0496	0.0209	0.0025	0.0858	0.0281	0.0017
	LATTICE	**0.0502**	**0.0275**	**0.0051**	**0.0744**	**0.0336**	**0.0038**	**0.1186**	**0.0425**	**0.0024**

SUMMARY

Accuracy results demonstrate that, with the only exception of LATTICE (whose trend is almost aligned with the existing literature) all other approaches' performance is heavily influenced by the hyper-parameter exploration and dataset characteristics. Indeed, even shallow models (e.g., VBPR) show competitive if not superior accuracy measures compared to more recent and complex solutions (e.g., MMGCN, GRCN).

Novelty and diversity (RQ2)

Datasets	Models	$k = 10$			$k = 20$			$k = 50$		
		EFD	Gini	iCov (%)	EFD	Gini	iCov (%)	EFD	Gini	iCov (%)
Office	VBPR	0.1753	0.3634	93.83	0.1479	0.396	10.23	0.1115	0.4413	99.59
	MMGCN	0.1140	0.0128	3.07	0.1027	0.0231	4.64	0.0845	0.0546	10.23
	MGAT	0.1079	0.0132	5.14	0.0963	0.0241	8.12	0.0792	0.0575	17.23
	GRCN	0.1215	**0.4587**	**99.01**	0.1051	**0.4892**	**99.79**	0.0829	**0.5286**	**100**
	LATTICE	**0.1827**	0.2128	87.86	**0.1513**	0.2652	95.90	**0.1125**	0.3414	99.30
Toys	VBPR	0.1948	0.2645	84.90	0.1527	0.3011	92.82	0.1051	0.3585	97.85
	MMGCN	0.0648	0.0989	37.87	0.0570	0.1450	52.51	0.0455	0.2296	72.88
	MGAT	0.0929	0.1036	40.95	0.0796	0.1439	55.71	0.0612	0.2183	76.24
	GRCN	0.1604	**0.3954**	**92.66**	0.1298	**0.4329**	**97.73**	0.0932	**0.4864**	**99.73**
	LATTICE	**0.2090**	0.1656	73.80	**0.1665**	0.2026	86.58	**0.1151**	0.2662	95.94
Clothing	VBPR	0.0502	0.2437	83.40	0.0413	0.2791	92.33	0.0291	0.3344	98.00
	MMGCN	0.0292	0.0136	7.58	0.0240	0.0236	12.44	0.0182	0.0493	23.34
	MGAT	0.0315	0.0201	11.05	0.0263	0.0326	17.36	0.0205	0.0622	30.90
	GRCN	0.0481	**0.3990**	**93.37**	0.0397	**0.4368**	**97.77**	0.0293	**0.4929**	**99.73**
	LATTICE	**0.0738**	0.1022	58.49	**0.0589**	0.1384	76.20	**0.0413**	0.2037	93.23

SUMMARY

While novelty results are almost aligned with the accuracy trends observed in RQ1, the diversity/coverage measures depict a different scenario. In this respect, GRCN seems to be the approach providing the most diversified item recommendations but at the expense of the accuracy, while VBPR manages to reach a more balanced performance among all metrics.

Politecnico di Bari

SISINF LAB

Conclusion and future work

Conclusion

- We show how a **careful hyper-parameter exploration** can lead **shallow** multimodal **approaches** (e.g., VBPR) to **be competitive** to more **recent** solutions.
- Other recent techniques such as **LATTICE** show to be **consistently outperforming** the other baselines.
- In terms of **novelty** and **diversity**, **GRCN** seems to be a **strong baseline** but **VBPR** is the solution reaching the most **balanced** accuracy, novelty, and diversity performance.

Future work

- Consider the **different impact** of **each modality** (**accepted** at the Workshop on Deep Multimodal Learning for Information Retrieval @ **ACM Multimedia 2023**).
- Additional **datasets** and **baselines**, **deeper** hyper-parameter **explorations**, recommendation metrics accounting for **bias** and **fairness**.

Politecnico di Bari

SISINF LAB

Thank you! Any questions?