The 29th ACM Conference on Multimedia

ACM multimedia
Chengdu, China OCT 20-24 2021

Mining Latent Structures for Multimedia Recommendation

Presented by Jinghao ZHANG

jinghao.zhang@cripac.ia.ac.cn

Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

CRIPAC 智能感知与计算研究中心 Center for Research on Intelligent Perception and Computing

Joint work with Yanqiao ZHU, Qiang LIU, Shu WU, Shuhui WANG, and Liang WANG

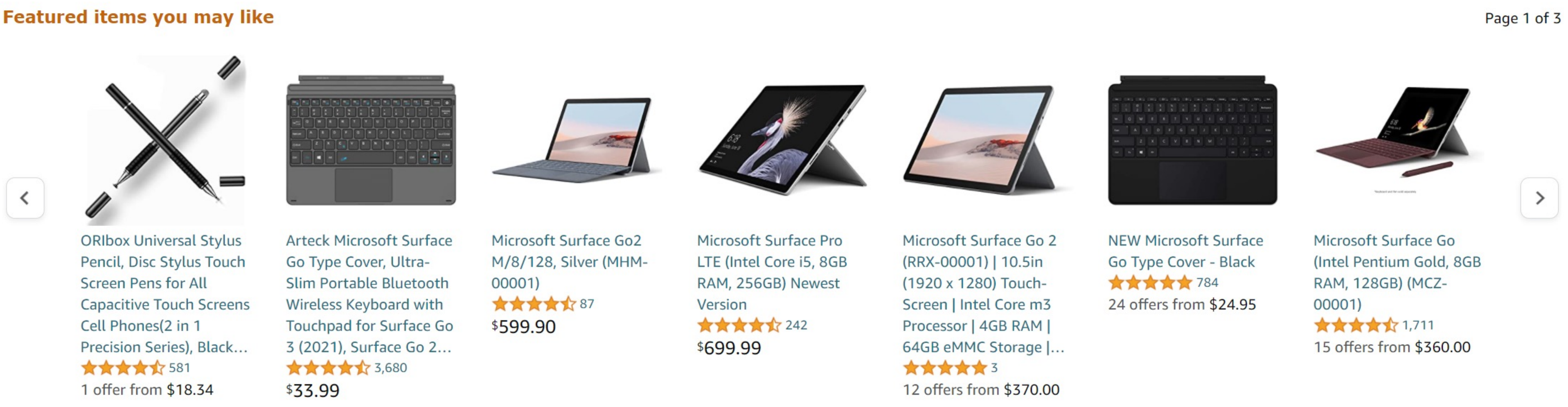
Outline

1. Preamble
2. The Proposed Method
3. Experiments
4. Concluding Remarks

Outline

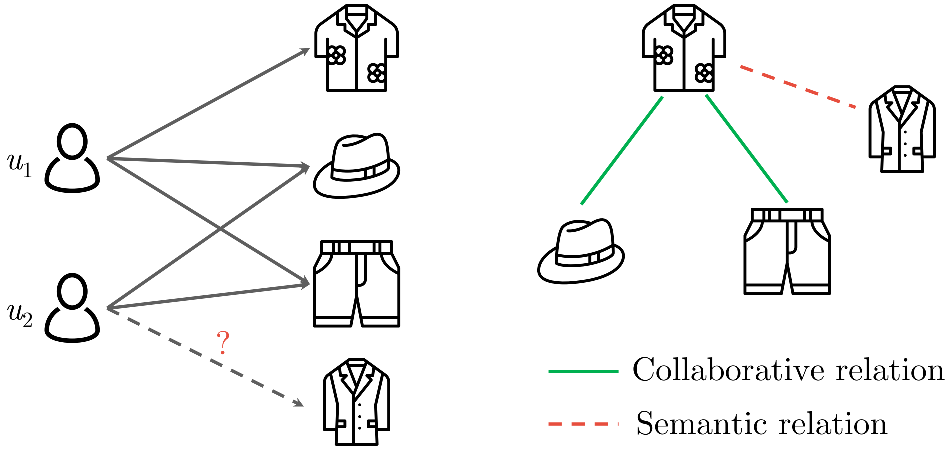
Background

- Multimedia content is of predominance in the modern Web era.
- Multimedia recommendation considers both **user-item interactions** and **item contents from various modalities** (e.g., visual, acoustic, and textual).



Motivation

- Previous work implicitly captures collaborative item-item relationships through high-order item-user-item relations.
- Considering that items are associated with rich contents in multiple modalities, we argue that the latent semantic item-item structures underlying these multimodal contents could be beneficial.

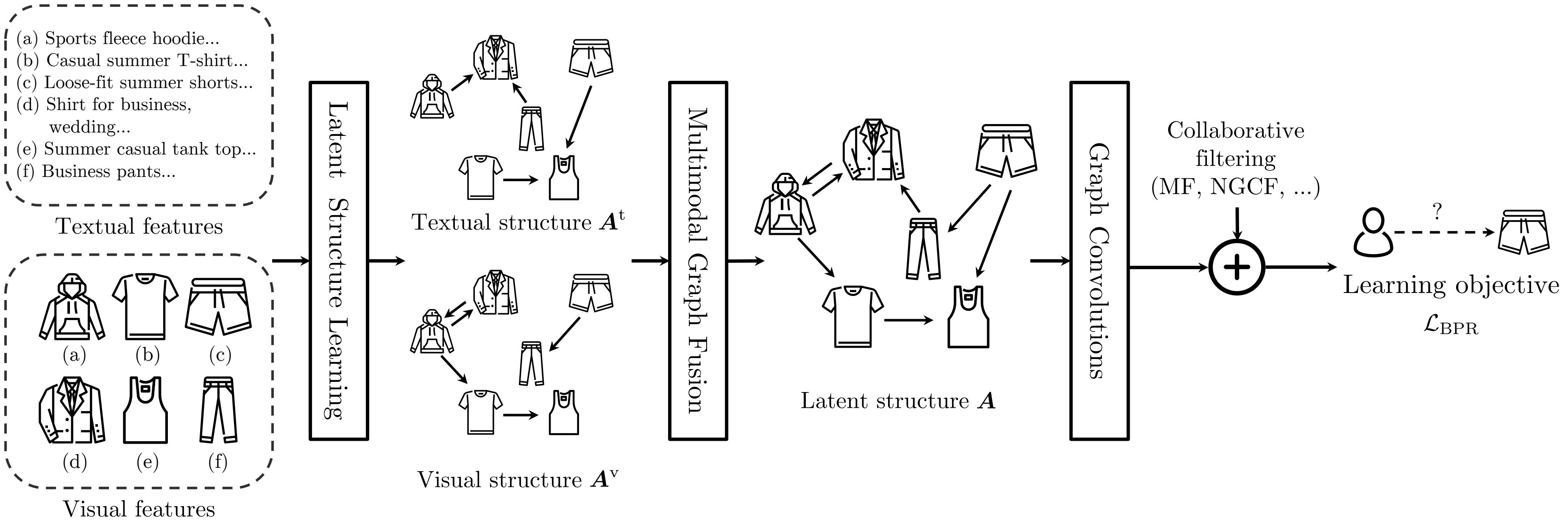


Outline

1. Preamble
2. **The Proposed Method**
3. Experiments
4. Concluding Remarks

The Proposed Framework

- Three main components in our LATTICE model:
  - Mining latent structures from multimodal features
  - Learning item representations with graph convolutions
  - Joint training with collaborative filtering objectives



Mining Latent Structures

- Construct initial kNN modality-aware graphs using raw multimodal features
  - Step 1. Quantify the semantic relationship by cosine similarity

$$\boldsymbol{S}_{ij}^{m} = \frac{(\boldsymbol{e}_i^m)^\top \boldsymbol{e}_j^m}{\|\boldsymbol{e}_i^m\| \|\boldsymbol{e}_j^m\|}$$

  - Step 2. Conduct kNN sparsification

$$\widehat{\boldsymbol{S}}_{ij}^{m} = \begin{cases} \boldsymbol{S}_{ij}^{m}, & \boldsymbol{S}_{ij}^{m} \in \text{top-}k(\boldsymbol{S}_i^m), \\ 0, & \text{otherwise.} \end{cases}$$

  - Step 3. Normalize the resulting adjacency matrix

$$\widetilde{\boldsymbol{S}}^{m} = (\boldsymbol{D}^m)^{-\frac{1}{2}} \widehat{\boldsymbol{S}}^{m} (\boldsymbol{D}^m)^{-\frac{1}{2}}$$

Mining Latent Structures

- Learn latent structures from transformed features
  - Step 4. Transform raw modality features into high-level features

$$\widetilde{\boldsymbol{e}}_i^m = \boldsymbol{W}_m \boldsymbol{e}_i^m + \boldsymbol{b}_m$$

  - Step 5. Repeat the above graph learning process using $\widetilde{\boldsymbol{e}}_i^m$
  - Step 6. Combine the learned graph $\widetilde{\boldsymbol{A}}^m$ with the initial graph $\widetilde{\boldsymbol{S}}^m$

$$\boldsymbol{A}^m = \lambda \widetilde{\boldsymbol{S}}^m + (1 - \lambda) \widetilde{\boldsymbol{A}}^m$$

  - Step 7. Aggregate modality-specific graphs in an adaptive way

$$\boldsymbol{A} = \sum_{m=0}^{|\mathcal{M}|} \alpha_m \boldsymbol{A}^m$$

Learning Item Representations

- After obtained item affinities, we perform simplified graph convolutional operations:

$$\boldsymbol{h}_i^{(l)} = \sum_{j \in \mathcal{N}(i)} \boldsymbol{A}_{ij} \boldsymbol{h}_j^{(l-1)}$$

  - We set the input item representation $\boldsymbol{h}_i^{(0)}$ as its corresponding ID embedding vector $\boldsymbol{x}_i$ rather than multimodal features.

Jointly Training with Downstream CF

- For any downstream collaborative filtering methods, we denote their user and item embeddings as $\widetilde{\boldsymbol{x}}_u$ and $\widetilde{\boldsymbol{x}}_i$.
- Then, we enhance item embeddings by adding high-order features learned through item graphs:

$$\widehat{\boldsymbol{x}}_i = \widetilde{\boldsymbol{x}}_i + \frac{\boldsymbol{h}_i^{(L)}}{\|\boldsymbol{h}_i^{(L)}\|_2}$$

- Finally, the user-item preference score is computed by:

$$\hat{y}_{ui} = \widetilde{\boldsymbol{x}}_u^\top \widehat{\boldsymbol{x}}_i$$

$$\mathcal{L}_{\mathrm{BPR}} = -\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}_u} \sum_{j \notin \mathcal{I}_u} \ln \sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right)$$

Outline

1. Preamble
2. The Proposed Method
3. **Experiments**
4. Concluding Remarks

Experimental Configurations

- Datasets

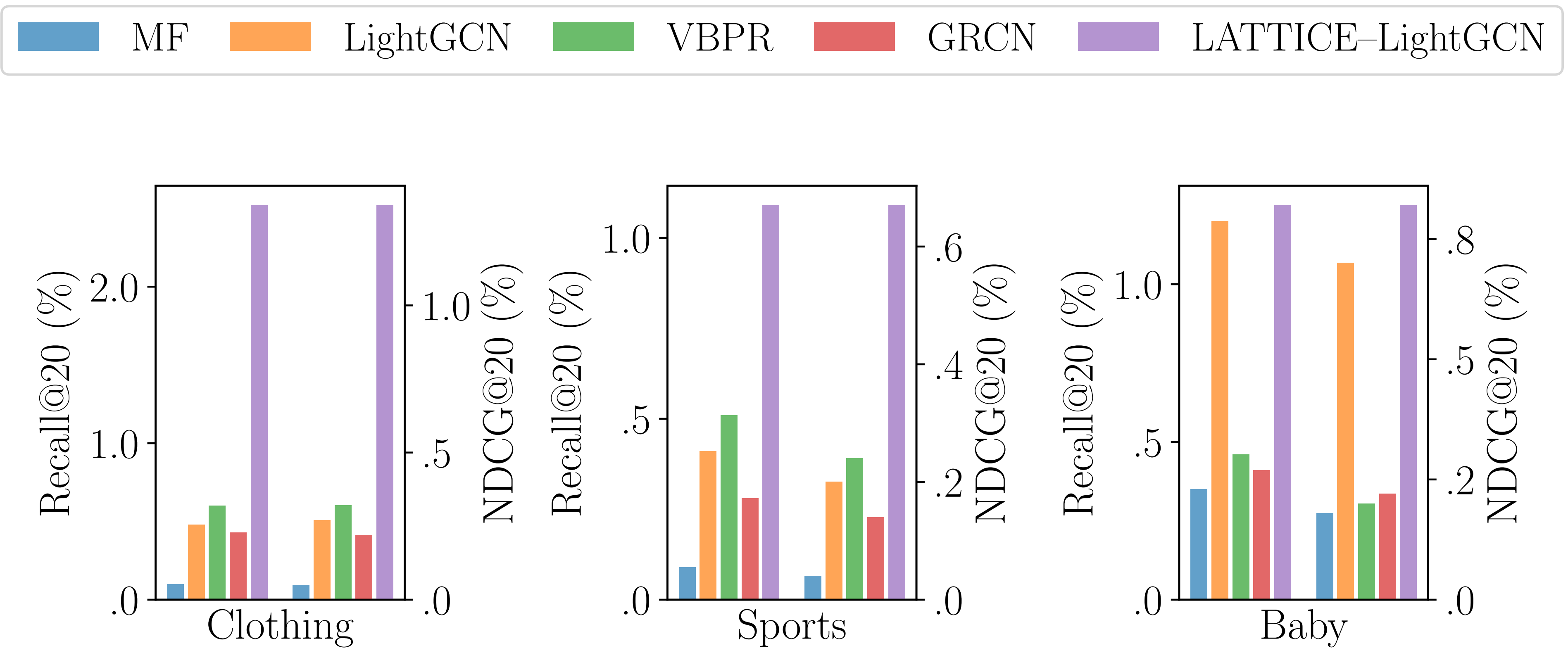
Dataset	# Users	# Items	# Interactions	Density
Clothing	39,387	23,033	237,488	0.00026
Sports	35,598	18,357	256,308	0.00039
Baby	19,445	7,050	139,110	0.00101

- Baselines:
  - Conventional CF models: MF, NGCF, and LightGCN
  - Content-aware recommenders: VBPR, MMGCN, and GRCN

Overall Performance

Model	Clothing			Sports			Baby		
	R@20	P@20	NDCG@20	R@20	P@20	NDCG@20	R@20	P@20	NDCG@20
MF	0.0191	0.0010	0.0088	0.0430	0.0023	0.0202	0.0440	0.0024	0.0200
NGCF	0.0387	0.0020	0.0168	0.0695	0.0037	0.0318	0.0591	0.0032	0.0261
LightGCN	0.0470	0.0024	0.0215	0.0782	0.0042	0.0369	0.0698	0.0037	0.0319
VBPR	0.0481	0.0024	0.0205	0.0582	0.0031	0.0265	0.0486	0.0026	0.0213
MMGCN	0.0501	0.0024	0.0221	0.0638	0.0034	0.0279	0.0640	0.0032	0.0284
GRCN	0.0631	0.0032	0.0276	0.0833	0.0044	0.0377	0.0754	0.0040	0.0336
LATTICE	**0.0710**	**0.0036**	**0.0316**	**0.0915**	**0.0048**	**0.0424**	**0.0829**	**0.0044**	**0.0368**
Improv.	12.5%	12.2%	14.6%	9.8%	8.7%	12.5%	9.9%	9.2%	9.5%

Cold-Start Performance



Ablation Studies

- CF+feats: Use multimodal features to replace the item representations learned from latent item graphs to combine with CF methods
- LATTICE/feats-CF: Use multimodal features to replace the item ID embedding as the input of GNNs

Model	Clothing			Sports			Baby		
	R@20	P@20	NDCG@20	R@20	P@20	NDCG@20	R@20	P@20	NDCG@20
MF	0.0191	0.0010	0.0088	0.0430	0.0023	0.0202	0.0440	0.0024	0.0200
MF+feats	0.0456	0.0023	0.0197	0.0674	0.0036	0.0304	0.0701	0.0037	0.0306
LATTICE/feats–MF	0.0519	0.0026	0.0224	0.0708	0.0038	0.0319	0.0729	0.0037	0.0326
LATTICE–MF	0.0577	0.0029	0.0246	0.0753	0.0040	0.0336	0.0767	0.0040	0.0339
Improv.	26.5%	25.9%	24.7%	11.7%	11.4%	10.7%	9.4%	9.4%	10.6%
NGCF	0.0387	0.0020	0.0168	0.0695	0.0037	0.0318	0.0591	0.0032	0.0261
NGCF+feats	0.0436	0.0022	0.0190	0.0748	0.0040	0.0344	0.0660	0.0035	0.0295
LATTICE/feats–NGCF	0.0480	0.0024	0.0212	0.0849	0.0043	0.0374	0.0713	0.0037	0.0307
LATTICE–NGCF	0.0488	0.0025	0.0216	0.0856	0.0045	0.0381	0.0727	0.0039	0.0313
Improv.	12.0%	11.9%	13.7%	14.5%	14.2%	10.9%	10.1%	9.4%	6.0%
LightGCN	0.0470	0.0024	0.0215	0.0782	0.0042	0.0369	0.0698	0.0037	0.0319
LightGCN+feats	0.0477	0.0024	0.0208	0.0754	0.0040	0.0350	0.0793	0.0042	0.0344
LATTICE/feats–LightGCN	0.0643	0.0033	0.0288	0.0832	0.0044	0.0386	0.0756	0.0040	0.0335
LATTICE–LightGCN	0.0710	0.0036	0.0316	0.0915	0.0048	0.0424	0.0836	0.0044	0.0373
Improv.	48.8%	48.4%	52.0%	21.3%	20.5%	21.3%	5.4%	5.2%	8.3%

Outline

1. Preamble
2. The Proposed Method
3. Experiments
4. **Concluding Remarks**

Concluding Remarks

- We highlight the importance of explicitly exploiting item relationships in multimedia recommendation, which supplement the collaborative signals modeled by traditional CF methods.
- We propose the latent structure mining method (LATTICE) for multimodal recommendation, which leverages graph structure learning to discover latent item relationships underlying multimodal features.
- Extensive experiments on three real-world datasets demonstrate the effectiveness of our proposed method.

THANKS

Code

Paper

Slides