Leveraging Content-Style Item Representation for Visual Recommendation

Yashar Deldjoo[1], Tommaso Di Noia[1], Daniele Malitesta[1], Felice Antonio Merra[2]*
[1]name.lastname@poliba.it, [2]felmerra@amazon.de
**corresponding author*

Politecnico di Bari

ECIR 2022

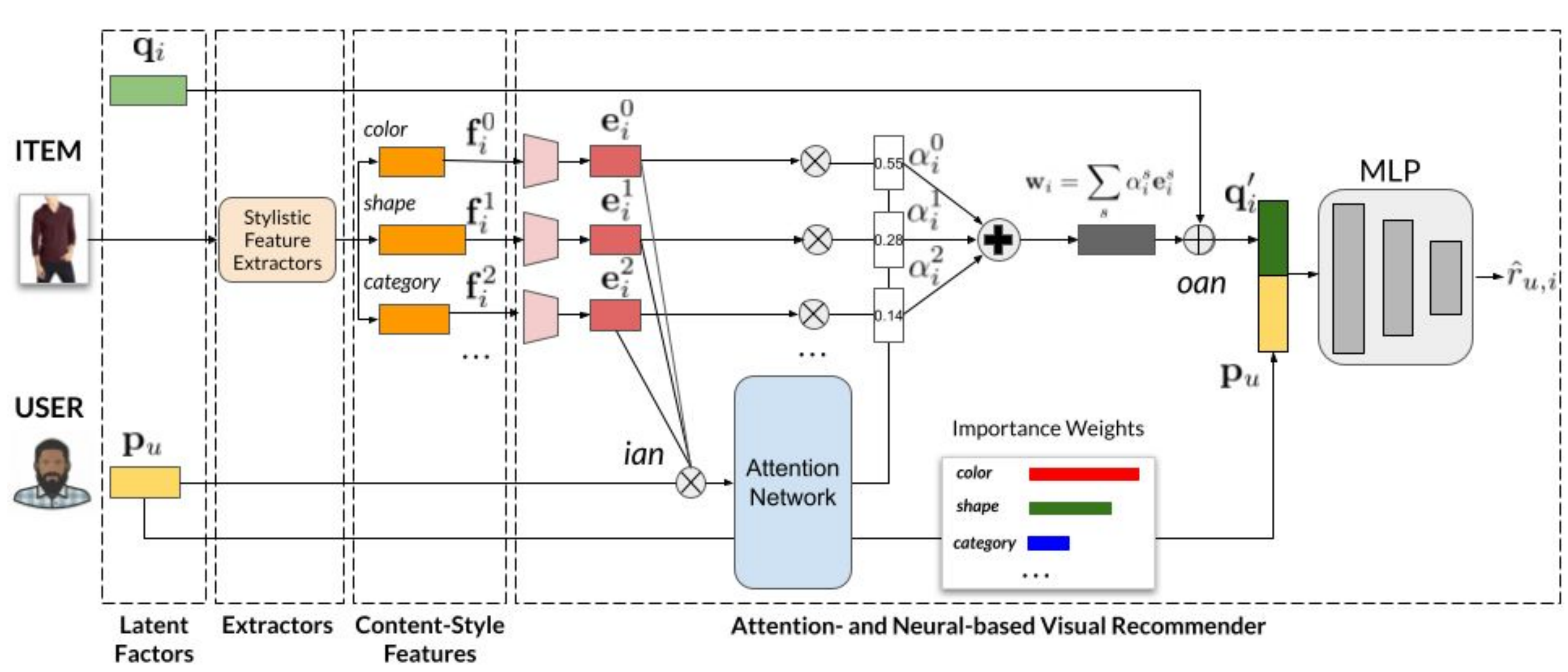
INTRODUCTION AND CONTRIBUTIONS

- Visual recommender systems [1, 2] (VRSs) use high-level visual features extracted from item images through pretrained convolutional neural networks [3, 4] as items' side information
- Recently [5, 6], attention mechanisms have been exploited to uncover users' visual attitude to finer-grained image characteristics on the content- [7] and region- [8] level
- However, the former requires side information (e.g., image tags or reviews) which may be difficult to collect, while the latter does not account for stylistic characteristics (e.g., color or texture)

Contributions

- We disentangle visual item representation on the stylistic content level (i.e., color, shape, and fashion item's category)
- We weight the importance of each feature on the user's visual preference through attention, and model user/item interactions through a neural architecture

METHODOLOGY



Let us consider:

- User $u \in \mathcal{U}$
- Item $i \in \mathcal{I}$
- Latent user factor $\mathbf{p}_u \in \mathbb{R}^{1 \times h}$
- Latent item factor $\mathbf{q}_i \in \mathbb{R}^{1 \times h}$
- Item content-style feature $\mathbf{f}_i^s \in \mathbb{R}^{1 \times v_s}$

We encode the item content-style feature: $\mathbf{e}_i^s = enc_s(\mathbf{f}_i^s)$

Then, we use a 2-layer attention network to weight the importance of each item content-style feature on the user:

$$a_{u,i}^s = \omega_2(\omega_1 ian(\mathbf{p}_u, \mathbf{e}_i^s) + \mathbf{b}_1) + \mathbf{b}_2 = \omega_2(\omega_1(\mathbf{p}_u \odot \mathbf{e}_i^s) + \mathbf{b}_1) + \mathbf{b}_2$$

and obtain a weighted stylistic representation of the item:

$$\mathbf{w}_i = \sum_{s \in \mathcal{S}} \alpha_{u,i}^s \mathbf{e}_i^s \quad \text{where} \sum_{s \in \mathcal{S}} a_{u,i}^s = 1$$

that we combine with the latent factor:

$$\mathbf{q}_i' = oan(\mathbf{q}_i, \mathbf{w}_i) = \mathbf{q}_i + \mathbf{w}_i$$

The final predicted rating is:

$$\hat{r}_{u,i} = out(concat(\mathbf{p}_u, \mathbf{q}_i'))$$

EXPERIMENTS AND RESULTS

Datasets

Dataset	Users	Items	Interactions	Density
Boys & Girls	1,425	5,019	9,213	0.00129
Men	16,278	31,750	113,106	0.00022

Results

RQ1) What are the accuracy and beyond-accuracy recommendation performance?

Model	*HR*	*nDCG*	*iCov*	*EFD*	*Gini*
Amazon Boys & Girls					
BPRMF	.01474	.00508	.68181	.00719	.28245
NeuMF	.02386	.00999	.00638	.01206	.00406
VBPR	.03018	.01287	.71030	.02049	.30532
DeepStyle	<u>.03719</u>	<u>.01543</u>	<u>.85017</u>	<u>.02624</u>	<u>.44770</u>
DVBPR	.00491	.00211	.00438	.00341	.00379
ACF	.01544	.00482	.70731	.00754	.40978
VNPR	.01053	.00429	.51584	.00739	.13664
Ours	**.03860**	**.01610**	**.89878**	**.02747**	**.49747**
Amazon Men					
BPRMF	<u>.01947</u>	<u>.00713</u>	.00605	.00982	.00982
NeuMF	.01333	.00444	.00076	.00633	.00060
VBPR	.01554	.00588	.59351	.01042	<u>.17935</u>
DeepStyle	.01634	.00654	**.84397**	**.01245**	**.33314**
DVBPR	.00123	.00036	.00088	.00069	.00065
ACF	.01548	.00729	.19380	.01147	.02956
VNPR	.00528	.00203	<u>.59443</u>	.00429	.16139
Ours	**.02021**	**.00750**	.28995	<u>.01242</u>	.06451

bold: best values, <u>underlined</u>: second-best values.

- On Boys & Girls, our solution and DeepStyle are the best and second-to-best models, and our approach outperforms the other baselines on novelty and diversity (e.g., *iCov* is around 90%)
- On Men, our solution is the most accurate one, even beating BPRMF, which covers only 0.6% of the catalogue (we reach 29% of the catalogue, and get the second best value on *EFD*)

RQ2) How performance is affected by different configurations of attention, *ian*, and *oan*?

Components		Boys & Girls		Men	
***ian*(·)**	***oan*(·)**	*HR*	*iCov*	*HR*	*iCov*
No Attention		.01263	.01136	.01462	.02208
Add	Add	.02316	.00757	.02083	.00076
Add	Mult	.02246	.00458	.00768	.00079
Concat	Add	.01404	.00518	**.02113**	.00076
Concat	Mult	.02456	.00458	.00891	.00085
Mult	*Add*	**.03860**	**.89878**	.02021	**.28995**
Mult	Mult	.02807	.00478	.01370	.01647

- All rows but *No Attention* lead to better-tailored recommendations
- The combination {Mult, Add} used in our approach is the most competitive on accuracy and beyond-accuracy metrics

Our proposed method reaches a competitive trade-off between accuracy and beyond-accuracy metrics

FUTURE WORK

- Extend the work to other visual recommendation scenarios (e.g., food and social media)
- Improve recommendation of extremely long-tail items, for which traditional CF is not beneficial

PAPER CODE:

github.com/sisinflab/Content-Style-VRSs

REFERENCES

[1] He and McAuley, The Web Conference, 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering.
[2] He and McAuley, AAAI, 2016. VBPR: visual bayesian personalized ranking from implicit feedback.
[3] Anelli et al., SIGIR, 2021. A study of defensive methods to protect visual recommendation against adversarial manipulation of images.
[4] Deldjoo et al., CVPR Workshops, 2021. A study on the relative importance of convolutional neural networks in visually-aware recommender systems.
[5] Chen et al., SIGIR, 2017. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention.
[6] Hou et al., IJCAI, 2019. Explainable fashion recommendation: A semantic attribute region guided approach.
[7] Cheng et al., ACM Trans. Inf. Syst., 2019. MMALFM: explainable recommendation by leveraging reviews and images.
[8] Wu et al., IEEE Access, 2020. Visual and textual jointly enhanced interpretable fashion recommendation.