A Study on the Relative Importance of Convolutional Neural Networks in Visually-Aware Recommender Systems

Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta*, Felice Antonio Merra
Politecnico di Bari, Italy
firstname.lastname@poliba.it

Abstract

Visually-aware recommender systems (VRSs) enhance the semantics of user-item interactions with visual features extracted from item images when they are available. Traditionally, VRSs leverage the representational power of pretrained convolutional neural networks (CNNs) to perform the item recommendation task. The adoption of CNNs is mainly attributed to their outstanding performance in representing visual data for supervised learning tasks, such as image classification. Their main drawback is that the learned representation of these networks is not entirely in line with the RS tasks — learning users' preferences.

This work aims to provide a better understanding of the representation power of pretrained CNNs commonly adopted by the community when integrated with state-of-the-art VRSs algorithms. In particular, we evaluate the recommendation performance of a suite of VRSs using several pretrained CNNs as the image feature extractors on two datasets from a real-world e-commerce platform. Additionally, we propose a novel qualitative and quantitative evaluation paradigm to assess the visual diversity of recommended items compared to the interacted user's items.

1. Introduction

With the increasing popularity of online services that provide users access to a wide range of services such as e-commerce (e.g., Zalando), multimedia content delivery (e.g., Netflix), and social networks (e.g., Instagram), the amount of available information has skyrocketed. Recommender systems (RSs) reduce the decision anxieties of over-choice by pointing users to a small set of items from a much larger set of items in the catalogue. Nowadays, RSs have grown to be an essential part of all large Internet retailers, making up to 35% of Amazon sales [25] or over 80% of the content watched on Netflix [7].

Recommendation based on user-item interactions, or collaborative filtering (CF) methods, has dominated the research in the RS community for years due to their superb recommendation quality. CF models infer users' preference on unseen items by leveraging the collaborative signal encoded in the recorded interactions (past behavioural data). However, in scenarios such as fashion [12], or food [11] recommendation, images associated with products can also impact the outcomes of users' decision making, as images attract attention, stimulate emotion, and shape users' first impression about products and brands. To extend the expressive power of RSs, visual-based recommender systems (VRSs) have emerged as an attempt to incorporate products' visual appearance of items into the design space of RS models [9]. The other advantage of VRS arises in cold-start situations, where new items added to the catalogue lack sufficient interactions, i.e., the so-called cold-items, thereby impeding the performance of CF models.

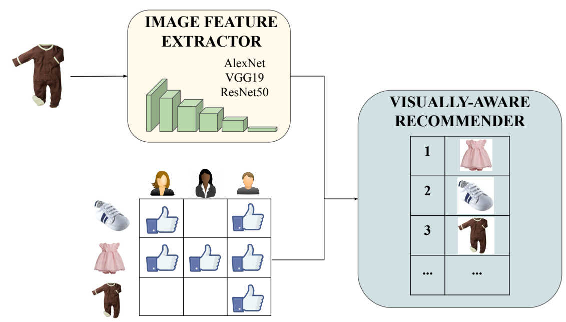

Figure 1: A Visually-Aware Recommender System (VRS).

Given the representational power of convolutional neural networks (CNNs) in capturing characteristics and semantics of the images in supervised learning tasks, such as image classification, state-of-the-art VRSs often exploit pretrained CNNs to implement the Image Feature Extractor (IFE) component of a VRS, as shown in Figure 1. This approach allows VRSs to exploit: *(i)* the high-level visual representational power of CNNs, and *(ii)* their ability to generalize on datasets different from the ones they were trained on, e.g., ImageNet [10]. Despite their success, there is a

*Authors are listed in alphabetical order. Corresponding author: Daniele Malitesta (daniele.malitesta@poliba.it).

lack of homogeneity in the selection of the pretrained networks in the literature, which usually happens to be a fixed choice. For instance, Hou et al. [17] propose an explainable fashion recommender system leveraging textual attributes, regions of item images, and a global visual profile of images extracted through AlexNet [21] , then Chen et al. [6] use VGG19 [32] to implement an explainable fashion recommender systems based upon image regions and user reviews, and finally Chen et al. [4] exploit ResNet50 [13] to generate a high-level description of recipe images which, along with textual descriptions, addresses the task of cross-modal recipe retrieval.

In this work, we aim at studying the impact of the three most popular pretrained CNN classes, namely AlexNet, VGG19, and ResNet50, used widely in the prior literature on a suite of competitive VRSs, contemplating four models, i.e., VBPR [16], DeepStyle [23], ACF [5], and VNPR [28]. The combinations of these CNNs and VRSs constitute the state-of-the-art for visual recommender models. Our contributions are two-fold: we evaluate to what extent different CNN architectural styles affect recommendation in terms of: *(i)* accuracy and beyond-accuracy metrics, and *(ii)* visual diversity of recommended items with respect to the ones previously consumed by each user.

2. Background and Related Work

Recommendation Problem. A recommendation problem seeks to find an automatic way to predict if —or to what extent— a user likes an unknown item through a *utility* function. Let $\mathcal{U}$ and $\mathcal{I}$ be the users and items sets, respectively. Given a utility function $g : \mathcal{U} \times \mathcal{I} \rightarrow \mathbb{R}$, we define a Recommendation Problem (RP) as $\forall u \in \mathcal{U}, i'_u = \arg\max_{i \in \mathcal{I}} g(u, i)$, with $i, i'_u \in \mathcal{I}$, where i'_u is an item not interacted by u yet. Furthermore, we set $R \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ as the user-item rating matrix (URM), where each element r_{ui} is either a continuous-valued rating assigned by user u to item i, i.e., explicit feedback, or a $0/1$-valued rating, i.e., implicit feedback. We refer to interacted items as positive and non-interacted ones as negative. Matrix Factorization (MF) [20], one of the most popular machine learning-driven approach for recommendation, maps each user (item) identifiers to a latent representation, i.e., $p_u \in \mathbb{P}^{1 \times h}$ $\left(q_i \in \mathbb{Q}^{1 \times h}\right)$, with $h << |\mathcal{U}|, |\mathcal{I}|$. The idea is to learn such embeddings to approximate URM through their dot product.

Visual Recommendation Problem. A Visual Recommendation Problem (VRP) tailors RP to the cases where item images are available, e.g., fashion and food recommendation. Let $\mathcal{X}$ be the set of item images. We aim at finding the image feature extraction function f to obtain the visual features of each image $f(x_i) = \varphi_i$, with $x_i \in \mathcal{X}$, enhancing, or even replacing, the recommendation-specific item representation. When pretrained CNNs are utilized as Image Feature Extractors (IFEs), it is common to extract the features on the layer activations, either convolutional or fully-connected.

Related Work on VRSs. Several works verified performance enhancements when integrating item visual features [31, 15, 16, 8]. The vast majority of them use high-level features extracted from CNNs, e.g., [15, 16], that could be either pretrained on a general-purpose dataset, e.g., ImageNet [10], or trained jointly with recommendation task, e.g., DVBPR [19]. As for the first category, VBPR [16] is the leading solution including visual features extracted from a pre-trained AlexNet [21] to extend the BPR-MF score function [30]. A year later, Liu et at. [23] proposed DeepStyle, a VBPR-based technique that assigns higher importance to the image style at the expense of the image category. Similarly, Niu et al. presented VNPR [28], which concatenates the PCA-reduced representation of item images extracted through an AlexNet-like architecture [38] to their recommendation embeddings before feeding it into a neural-based recommender model. Then, Chen et al. [5] implemented ACF, which —differently from the previous approaches— adopts the feature maps extracted from a convolutional layer of a pretrained ResNet152 [13] to weight the different regions within users' positive item images through attention mechanisms. Chen et al. [6] designed an attention-based approach for explainable fashion recommendations by exploiting a pretrained VGG19 [13].

While big efforts have been dedicated to building accurate VRSs, we noticed a lack in exploring how much the chosen pretrained CNN would impact on the recommendation performance. Indeed, we found that AlexNet, ResNet, and VGG are the most popular networks, i.e., at least 7 papers for the first [26, 15, 16, 14, 23, 28, 17], 6 for the second [5, 37, 4, 29, 33, 3], and 3 for the third one [6, 36, 35], but there are no exhaustive studies to verify their differences. In this work, we aim to fill this gap by studying various configurations of state-of-the-art VRSs using standard pretrained CNNs, i.e., AlexNet, VGG19, and ResNet50.

3. Experiment Settings

3.1. Datasets

We investigate two fashion datasets, i.e., `Amazon Baby` and `Amazon Boys & Girls` [15, 26]. Both were filtered through the 5-core technique as suggested in [15, 16] to avoid cold-start users, thus resulting in the following statistics: the former counts 606 users, 1761 items, and 3882 registered interactions, while the latter covers 600 users and 2760 items, with 3910 ratings.

3.2. Image Feature Extractors

We study three IFEs: AlexNet, VGG19, and ResNet50. The first, **AlexNet** [21], is a 8-layer CNN, i.e., 5 convolutional and 3 fully-connected layers. This is one of the first architectures to introduce `ReLU` activation function [27] to address the saturation issue of the `tanh` function. The second, **VGG19** [32], is one of the first *deep*-CNN, consisting of 19 layers, i.e., 16 convolutional and 3 fully-connected layers. All convolutions are built on a 3×3 kernel, and, like AlexNet, `ReLU` is the activation function. The last, **ResNet50** [13], is the 50-deep CNN belonging to the ResNet family. It adopts residual blocks to tackle the training degradation problem observed in deep-CNNs. The ResNet family won the ILSVRC-2015 [24], outperforming their non-residual counterparts, e.g., VGG19.

3.3. Visual-based Recommender Models

We explore four VRSs: VBPR, DeepStyle, VNPR, and ACF. The first, Visual Bayesian Personalized Ranking (**VBPR**) [16], calculates the predicted rating for a user u and an item i as $\hat{r}_{ui} = p_u^T q_i + \theta_u^T \mathbf{E} \varphi_i$, where θ_u is the user's *visual* latent vector, φ_i is the item feature extracted from a fully-connected layer, and $\mathbf{E}$ is an embedding matrix to project φ_i into θ_u's spacn. **DeepStyle** [23], updates the VBPR score function by subtracting a $p_u^T c_i$ term where c_i embodies the categorical information of i. Visual Neural Personalized Ranking (**VNPR**) [28], computes the (u, i) preference score with a MLP whose input is the concatenation of the element-wise product of (p_u, q_i) and $(v_u, \hat{\varphi}_i)$, where the latter consists of the visual user profile and the PCA compression of φ_i . Attentive Collaborative Filtering (**ACF**) [5], predicts the user's score of an unrated item using two attention networks to weigh its importance in the set of u-positive items and the regions within these images. The ACF feature is the feature map extracted from a convolutional layer.

3.4. Evaluation Metrics

We study accuracy and beyond-accuracy metrics evaluated on top-k recommendation lists. As for the *accuracy* measures, we adopt the *recall* ($\mathbf{Rec}@k$) —the fraction of recommended products in the top-k that hit test items— and the *area under the ROC curve* ($\mathbf{AUC}$) —a k-independent metric defined as the probability of ranking a positive item more than a random negative one. Then, the *beyond-accuracy* measures are the *ratio of coveraged items* ($\mathbf{iCov}@k$) —the percentage of recommended items in the top-k lists— and the *expected free discovery* ($\mathbf{EFD}@k$) —a measure of the model capacity of suggesting relevant long-tail (unpopular) items [34]. All the above cited metrics range from 0 to 1, the closer to 1 the better.

3.5. Reproducibility

We split the datasets by adopting the temporal leave-one-out paradigm, i.e., for each user, the test and validation sets contain the last and second-to-last interactions. We apply a grid-search to tune the hyperparameters on the validation set. We release our code[1] implemented in `Elliot` [2].

Table 1: Recommendation results on top-100 lists.

Dataset	VRS	IFE	Rec	AUC	iCov	EFD
`Amazon Baby`	VBPR	AlexNet	.1304	.6308	.9886	.0142
		VGG19	.1568	.6344	.9875	.0162
		ResNet50	**.2063**	**.6475**	**.9915**	**.0246**
	DeepStyle	AlexNet	.1337	.6094	.9994	.0155
		VGG19	.1683	.6372	.9960	.0191
		ResNet50	**.2195**	**.6400**	**1.000**	**.0271**
	ACF	AlexNet	**.1271**	**.5544**	**.7910**	**.0158**
		VGG19	.1073	.5477	.7763	.0132
		ResNet50	.1023	.5532	.7791	.0122
	VNPR	AlexNet	.0561	.5221	.6303	.0061
		VGG19	.0891	.5349	.8001	.0111
		ResNet50	**.1221**	**.5817**	**.9733**	**.0141**
`Amazon Boys & Girls`	VBPR	AlexNet	.1033	.6348	.9808	.0137
		VGG19	.1133	.6262	.9681	.0140
		ResNet50	**.1250**	**.6606**	**.9837**	**.0146**
	DeepStyle	AlexNet	.0983	.6160	.9993	.0114
		VGG19	.1133	.6307	**.9996**	**.0168**
		ResNet50	**.1250**	**.6402**	.9957	.0152
	ACF	AlexNet	**.0450**	.5120	**.8043**	.0047
		VGG19	.0433	.4955	.7424	**.0049**
		ResNet50	.0300	**.5235**	.7518	.0029
	VNPR	AlexNet	.0317	.5018	.5319	.0043
		VGG19	.0417	.5358	.6272	.0051
		ResNet50	**.0800**	**.5727**	**.9667**	**.0094**

4. Results and Discussion

This section evaluates the effects of varying the IFE on the top of the tested VRSs. All the metrics are computed for the top-100 recommendations. We will refer to each of them without the k term, e.g., **iCov** instead of **iCov**@100.

Analysis of Recommendation Results. Table 1 reports the accuracy and beyond-accuracy recommendation metrics. To begin with, it can be observed that VRSs built upon ResNet50 exhibit the best recommendation performance. Indeed, we notice that the VRS variants adopting visual features extracted from ResNet50 outperform the other IFE in 72% of the experimented cases. AlexNet settles as the second quality-level IFE, leaving VGG19 to the last place despite its widely-recognized ability to extract visual and stylistic content from images [18]. We may explain this, saying that deeper convolutional networks with residual blocks, such as ResNet50, produce more accurate recommendations thanks to their representational power.

Additionally, we observe that the positive impact of ResNet50 on recommendation is uniformly not confirmed for ACF. In this setting, AlexNet is the pre-trained CNN that ensures the best accuracy performance in both the tested datasets. For instance, ACF using AlexNet features has a **Rec** equal to 0.0450, compared to the ResNet50 value of 0.0300. The reason for these outcomes could lie in the specific model characteristic. Indeed, differently from the other explored VRSs which take the output of a *fully-connected* layer as input, ACF leverages visual features extracted from a *convolutional* layer for the sake of the component-level

[1] https://github.com/sisinflab/CNNs-in-VRSs

Table 2: Average *visual diversity* (**VisDiv**) on top-100 lists.

Dataset	**VRS**	**IFE**		
		AlexNet	VGG19	ResNet50
Amazon Baby	VBPR	13.16	14.92	**17.05**
	DeepStyle	14.52	14.10	**16.64**
	ACF*	53.93	**59.93**	52.27
	VNPR	7.40	**20.75**	10.32
Amazon Boys & Girls	VBPR	10.16	15.62	**20.67**
	DeepStyle	12.32	14.27	**20.08**
	ACF*	58.46	**70.73**	48.31
	VNPR	11.96	8.98	**27.27**

* *Visual features have been flattened for t-SNE.*

attention mentioned in Section 3.3. As convolutional layers catch a lower-level representation of images compared to fully-connected ones, it entails that the different extraction layer is dramatically reducing the observed importance of IFE's depth in VRSs.

Furthermore, we evaluate the effects of varying the IFE on beyond-accuracy metrics, i.e., **iCov** and **EFD**. Similarly to the analysis of the accuracy-based results, both the beyond-accuracy measures reach the best values when ResNet50 is used as IFE. For example, considering the **EFD** measured for DeepStyle on Amazon Baby, the usage of ResNet50 produces the best metric value, i.e., 0.0271. In this setting, it is interesting to notice that only by changing the IFE from the original paper [23], i.e., AlexNet, we obtain an **EFD** improvement of +75%. This novel finding could be explained by the fact that the extracted features of deeper and complex CNNs, like ResNet50, allow learning more diverse users' preferences.

In summary, the results validate the hypothesis according to which the impact strength on VRSs can significantly vary based on the pretrained CNN employed. In fact, the deeper networks, such as ResNet50, seem to provide a much higher quality of recommendation in strong VRSs such as DeepStyle and VBPR. For average-quality VRSs, not a single CNN type outperforms the rest. Finally, we witness the same trend on beyond-accuracy metrics, such as item coverage and novelty, which directly measure the impact on users, platform owners, and third-party sellers in terms of economic gains and experience satisfaction [1, 22].

Analysis of Users Visual Profile. This section quantitatively and qualitatively evaluates to what extent each user's top-100 recommended items are visually similar, or dissimilar, to the list of positive ones. To address this analysis, we define the *visual diversity* (**VisDiv**@k) as the Euclidean distance between the visual features centroids extracted from both the positive and top-100 recommended items. Such distance is calculated after the application of the *t-SNE* algorithm to the feature embeddings to project them into a 2D latent space, which also come in handy for visualization purposes (see later).

Table 2 reports the average **VisDiv** on all users. Investigating this quantitative metric, it can be observed that the settings with higher **VisDiv** are connected to the ones with the most *accurate* and *diverse* recommendation per-

(a) **AlexNet** (b) **ResNet50**

Figure 2: Positive (green) and top-5 (red) item features in the latent space for (a) AlexNet and (b) ResNet50. The **VisDiv**@5 (the line connecting the two centroids) are 67.83 and 416.22 respectively.

formance in Table 1. For instance, when comparing VBPR experiments varying the IFE, both visual and recommendation metrics reach the highest values when using ResNet50. To be specific, **VisDiv**, i.e., 17.05 and 20.67, **Rec**, i.e., 0.2063 and 0.1250, **EFD**, i.e., 0.0246 and 0.0146, on Amazon Baby and Amazon Boys & Girls respectively, confirm that a higher **VisDiv** value can be linked to better recommendation performance. Coherently, comparing the bold values of Table 1 and Table 2, it can be seen that VRSs using the IFE with ResNet50 produce the best performing and most visually-diverse recommendations.

To conclude, Figure 2 helps to inspect the visual differences of the positive and top-5 VBPR-based recommended items of a user sampled from Amazon Boys & Girls when the image features are extracted from AlexNet (Figure 2a) and ResNet50 (Figure 2b). It can be observed that, while the usage of AlexNet leads to the recommendation of items visually similar to the positive ones, i.e., all items are in the "trekking shoes" category as shown in Figure 2a, the application of ResNet50 makes recommendations more diverse, i.e., boots and socks in Figure 2b, and even with variable colour, e.g., the recommended jackets.

5. Conclusion and Future Work

In this work, we investigated the effect of choosing the CNN model on top of a VRS to extract item images' visual features. We performed 24 experimental combinations varying VRSs, CNNs used as IFE, and datasets. We proved that a deeper IFE, i.e., ResNet50, ensures high accuracy and beyond-accuracy recommendation performance. Moreover, ResNet50 has shown quantitatively and qualitatively to produce the most diverse recommended products under both a recommendation and visual-appearance perspectives. We plan to extend this study with other popular CNNs, e.g., Inception and EfficientNet, and domain-specific ones, e.g., DeepFashion, but also VRSs involving an end-to-end training, e.g., DVBPR, on larger datasets.

Acknowledgment. The authors acknowledge partial support of the projects: Servizi Locali 2.0, PON ARS01_00876 Bio-D, PON ARS01_00821 FLET4.0, PON ARS01_00917 OK-INSAID, H2020 PASSPARTOUT.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.*, 30(1):127–158, 2020. 4

[2] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco M. Donini, and Tommaso Di Noia. Elliot: a comprehensive and rigorous framework for reproducible recommender systems evaluation. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Online - July 11 - 15, 2021*. ACM, 2021. 3

[3] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. A study of defensive methods to protect visual recommendation against adversarial manipulation of images. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Online - July 11 - 15, 2021*. ACM, 2021. 2

[4] Jingjing Chen, Chong-Wah Ngo, Fuli Feng, and Tat-Seng Chua. Deep understanding of cooking procedure for cross-modal recipe retrieval. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1020–1028. ACM, 2018. 2

[5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 335–344. ACM, 2017. 2, 3

[6] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 765–774. ACM, 2019. 2

[7] Sameer Chhabra. Netflix says 80 percent of watched content is based on algorithmic recommendations. https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/. Accessed: 2021-03-13. 1

[8] Xiaoya Chong, Qing Li, Howard Leung, Qianhui Men, and Xianjin Chao. Hierarchical visual-aware minimax ranking based on co-purchase data for personalized recommendation. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2563–2569. ACM / IW3C2, 2020. 2

[9] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. Recommender systems leveraging multimedia content. *ACM Comput. Surv.*, 53(5):106:1–106:38, 2020. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 1, 2

[11] David Elsweiler, Christoph Trattner, and Morgan Harvey. Exploiting food choice biases for healthier recipe recommendation. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 575–584. ACM, 2017. 1

[12] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5337–5345. Computer Vision Foundation / IEEE, 2019. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 3

[14] Ruining He, Chunbin Lin, Jianguo Wang, and Julian J. McAuley. Sherlock: Sparse hierarchical embeddings

for visually-aware one-class collaborative filtering. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3740–3746. IJCAI/AAAI Press, 2016. 2

[15] Ruining He and Julian J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM, 2016. 2

[16] Ruining He and Julian J. McAuley. VBPR: visual bayesian personalized ranking from implicit feedback. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 144–150. AAAI Press, 2016. 2, 3

[17] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W. Zheng, and Qi Liu. Explainable fashion recommendation: A semantic attribute region guided approach. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4681–4688. ijcai.org, 2019. 2

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711. Springer, 2016. 3

[19] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. Visually-aware fashion recommendation and design with generative image models. In Vijay Raghavan, Srinivas Aluru, George Karypis, Lucio Miele, and Xindong Wu, editors, *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 207–216. IEEE Computer Society, 2017. 2

[20] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. 2

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. 2

[22] Mounia Lalmas. Personalising and diversifying the listening experience. In Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich, editors, *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, page 3. ACM, 2020. 4

[23] Qiang Liu, Shu Wu, and Liang Wang. Deepstyle: Learning user preferences for visual recommendation. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 841–844. ACM, 2017. 2, 3, 4

[24] Wei Liu, Olga Russakovsky, Jia Deng, Fei-Fei Li, and Alex Berg. Imagenet large scale visual recognition challenge 2015. `http://www.image-net.org/challenges/LSVRC/2015/`. Accessed: 2021-03-13. 3

[25] Ian MacKenzie, Chris Meyer, and Steve Noble. How retailers can keep up with consumers. *McKinsey & Company*, 18, 2013. 1

[26] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 43–52. ACM, 2015. 2

[27] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress, 2010. 3

[28] Wei Niu, James Caverlee, and Haokai Lu. Neural personalized ranking for image recommendation. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 423–431. ACM, 2018. 2, 3

[29] Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. Taamr: Targeted adversarial attack against multimedia recommender systems. In *50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN Workshops 2020, Valencia, Spain, June 29 - July 2, 2020*, pages 1–8. IEEE, 2020. 2

[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. In Jeff A. Bilmes and Andrew Y. Ng, editors, *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 452–461. AUAI Press, 2009. 2

[31] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 869–877. IEEE Computer Society, 2015. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2, 3

[33] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. Adversarial training towards robust multimedia recommender system. *IEEE Trans. Knowl. Data Eng.*, 32(5):855–867, 2020. 2

[34] Saúl Vargas. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, page 1281. ACM, 2014. 3

[35] Le Wu, Lei Chen, Richang Hong, Yanjie Fu, Xing Xie, and Meng Wang. A hierarchical attention model for social contextual image recommendation. *IEEE Trans. Knowl. Data Eng.*, 32(10):1854–1867, 2020. 2

[36] Qianqian Wu, Pengpeng Zhao, and Zhiming Cui. Visual and textual jointly enhanced interpretable fashion recommendation. *IEEE Access*, 8:68736–68746, 2020. 2

[37] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. Interpretable fashion matching with rich attributes. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 775–784. ACM, 2019. 2

[38] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 487–495, 2014. 2