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.

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
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 $U$ and $I$ be the users and items sets, respectively. Given a utility function $g : U \times I \to \mathbb{R}$, we define a Recommendation Problem (RP) as $\forall u \in U$, $i' \in I$, $i'' \in I$, where $i' \neq i''$ is an item not interacted by $u$. Furthermore, we set $R \in \mathbb{R}^{U \times I \times [I]}$ 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., $p_u \in \mathbb{R}^{1 \times h}$ ($q_i \in \mathbb{R}^{1 \times h}$), with $h << |U|$, $|I|$. 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 $X$ be the set of item images. We aim at finding the image feature extraction function $\phi$ to obtain the visual features of each image $\phi(x_i) = \varphi_i$, with $x_i \in X$, 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 al. [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 ResNet52 [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 VRS 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 func-
tion [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 \( \hat{r}_{ui} = p_{ui}^T q_i + \theta_u \phi_i \), where \( \theta_u \) is the user’s visual latent vector, \( \phi_i \) is the item feature extracted from a fully-connected layer, and \( E \) is an embedding matrix to project \( \phi_i \) into \( \theta_u \)’s space. DeepStyle [23], updates the VBPR score function by subtracting a \( p_{ui}^T c_i \) 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 \( (v_u, \phi_i) \), where the latter consists of the visual user profile and the PCA compression of \( \phi_i \). 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 (Rec@\( k \)) —the fraction of recommended products in the top-\( k \) that hit test items— and the area under the ROC curve (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 covered items (iCov@\( k \)) —the percentage of recommended items in the top-\( k \) lists— and the expected free discovery (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 code¹ implemented in Elliot [2].

### 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

---

¹https://github.com/sisinflab/CNNs-in-VRSs
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 +7.5%. 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 performance 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


