

# Adversarial Attacks against Visual Recommendation: an Investigation on the Influence of Items' Popularity

Vito Walter Anelli<sup>1</sup>, Tommaso Di Noia<sup>1</sup>, Eugenio Di Sciascio<sup>1</sup>, Daniele Malitesta<sup>1</sup> and Felice Antonio Merra<sup>1,2</sup>

<sup>1</sup>Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy

<sup>2</sup>The authors are in alphabetical order. Corresponding author: Felice Antonio Merra ([felice.merra@poliba.it](mailto:felice.merra@poliba.it)).

## Abstract

Visually-aware recommender systems (VRSs) integrate products' image features with historical users' feedback to enhance recommendation performance. Such models have shown to be very effective in different domains, ranging from fashion, food, to point-of-interest. However, test-time adversarial attack strategies have recently unveiled severe security issues on these recommender models. Indeed, adversaries can harm the integrity of recommenders by uploading item images with human-imperceptible adversarial perturbations capable of pushing a target item into higher recommendation positions. Given the importance of items' popularity on the recommendation performance, in this work, we evaluate whether there is an influence of items' popularity on the attacks' effectiveness. To this end, we perform three state-of-the-art adversarial attacks against VBPR (a standard VRS) by varying the adversary knowledge (white- vs. black- box) and capability (the magnitude of the perturbation). The results obtained evaluating attacks on two real-world datasets shed light on the remarkable efficacy of the attacks against the least popular items' when planning novel defenses.

## Keywords

Adversarial Machine Learning, Visual Recommender Systems, Collaborative Filtering

## 1. Introduction

Recommender systems (RSs) try to unveil the hidden relationships among users and items on popular e-commerce platforms (e.g., Amazon, Zalando) by presenting personalized lists of recommendations, thus supporting customers in the decision-making process. When the user's visual taste matters, in scenarios such as fashion [1], food [2], or point-of-interest [3] recommendations, visually-aware recommender systems (VRSs) have recently proven to provide superior results by leveraging the representational power of (pretrained) convolutional neural networks (CNNs) to extract meaningful item visual representations and inject them into the preference learning process to model the users' visual attitude towards products [4, 5, 6, 7]. For instance, He and McAuley [4] proposed VBPR, a popular matrix factorization (MF)-based VRS that integrates visual features extracted from a pre-trained CNN (i.e., AlexNet [8]).

While transferring the visual knowledge of pretrained CNNs on the recommendation task


---

OHARS'21: Second Workshop on Online Misinformation- and Harm-Aware Recommender Systems, October 2, 2021, Amsterdam, Netherlands

✉ [vitowalter.anelli@poliba.it](mailto:vitowalter.anelli@poliba.it) (V. W. Anelli); [tommaso.dinoia@poliba.it](mailto:tommaso.dinoia@poliba.it) (T. Di Noia); [eugenio.disciascio@poliba.it](mailto:eugenio.disciascio@poliba.it) (E. Di Sciascio); [daniele.malitesta@poliba.it](mailto:daniele.malitesta@poliba.it) (D. Malitesta); [felice.merra@poliba.it](mailto:felice.merra@poliba.it) (F. A. Merra)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

has represented a turning-point in the RecSys community, few have already considered the collateral and negative impact of adversarial attacks against deep/convolutional neural networks (DNNs/CNNs) used in visually-aware recommendations [9]. To date, there exists a plethora of adversarial attack strategies in the computer vision domain whose purpose is to perturb images and mislead the classification performance imperceptibly. In this set, FGSM [10], PGD [11], and Carlini & Wagner [12] represent the milestones in adversarial machine learning (AML). In collaborative filtering recommendations, He et al. [13] have proposed and demonstrated the efficacy of adversarial perturbation of MF model embeddings in corrupting the model performance. Then, they have designed an adversarial training method to robustify the model performance against the previously proposed perturbations. Their experimental flow has also been applied in [14]. Here, the authors have tested the first adversarial procedures (attacks/defenses) in a visually-aware recommendation model. Indeed, they have attacked, and later defended, VBPR against adversarial perturbations applied on the visual embeddings extracted from a pretrained CNN (i.e., ResNet50 [15]). While Tang et al. [14] have worked on feature-level perturbations, later Di Noia et al. [16], Anelli et al. [17] have studied and designed the first set of targeted adversarial attack methods to be directly performed against input product images (and not the visual features) to increase the recommendation probability of low-recommended categories of products by **poisoning** the training data with their adversarial samples.

More recently, Liu and Larson [18] and Cohen et al. [19] have proposed novel adversarial attack procedures that perturbs product images to push/nuke an item during the inference/testing phase (i.e., **evasive attacks**). Both works have released *black-box* and *white-box* adversarial methods where, in the first scenario, the adversary does not know the recommender model, while in the second, the attacker has complete access to the model, input, and output. However, both training and testing time attacks have been evaluated considering their efficacy on pushing the target/victim items into top- $K$  recommendation lists or increasing the preference scores without taking into account the different levels of items' popularity (i.e., the number of interactions recorded on each item in the training set). Indeed, considering the influence of different levels of item's popularity on the recommendation performance [20, 21, 22, 23], we found a lack of investigation on their potential effects on the efficacy of adversarial attacks.

In this work, motivated by the previous observations, we explore the performance of VBPR [4], a pioneering VRS, under test-time attacks. In particular, we investigate both *black-box* and *white-box* settings, and we split the target items into four groups based on their popularity to understand whether there could be a connection between attacks efficacy and the number of feedbacks received by the target item. Our contributions may be summarized as follows:

- we provide an extensive evaluation of three state-of-the-art adversarial attacks against visual-based recommendations in multiple settings, varying the adversary knowledge (i.e., black- and white-box), the adversarial capability (i.e., the maximum variation of each image pixel, that is  $\epsilon \in \{4, 8, 16\}$ ); and evaluating their performance on for groups of target items (i.e., Low Popular, Mid-Low Popular, Mid-High Popular, and High Popular);
- we measure and discuss the changes in the preference scores predicted from the trained VRS according to the variations of the predicted preference scores and the fraction of times a target item has received a preference score higher than the one before the attack;

- we investigate and compare the effectiveness in pushing the target items (divided again by popularity groups) in the top- $K$  position of the model generated recommendation lists.

We conduct experiments on two datasets from the Amazon domain [24, 25] to validate the effectiveness of the proposed model for the task of personalized visual recommendation.

## 2. Related Work

Recommender Systems (RSs) may rely on additional side information (e.g., images, audio, and text) to enhance the item representation and provide more tailored recommendations. Indeed, in domains such as fashion [1], food [2], and point-of-interest [3], product images displayed on online platforms can positively drive the final users’ decision. In this respect, to extend the expressive power of RSs, visually-aware recommender systems (VRSs) have recently proposed to incorporate products’ visual appearance of items into recommendation [4, 5, 6, 7, 26]. Given the representational power of convolutional neural networks (CNNs) in capturing high-level images’ characteristics, state-of-the-art VRSs often integrate visual features extracted via a CNN —pre-trained, e.g., [4, 6, 27, 28], or learned end-to-end, e.g., [29, 30]. When it comes to adversarially attacking VRSs, the literature recognizes two main approaches, by perturbing the visual appearance of items either on the *feature*-level (i.e., the visual embeddings extracted from the CNNs) or on the *image*-level (i.e., the item images). On the one side, the *feature*-level attacks, Tang et al. [14] designed and implemented a framework to robustify the model by He and McAuley [4] against visual feature perturbations by leveraging adversarial training, while Paul et al. [31] proposed an aesthetic-based VRS which is adversarially defended by adopting an iterative adversarial training procedure on the aesthetic features. On the other hand, the *image*-level attacks, Di Noia et al. [16], Anelli et al. [17] run state-of-the-art adversarial attacks in computer vision for poisoning the dataset with adversarial product images to alter the training of the model in order to push the target items towards higher recommendation positions. In addition, Liu and Larson [18], Cohen et al. [19] perturbed the item images at test-time in the realistic scenario that an adversary can upload an altered version of a product picture after the training of the model. In this work, we only consider *image*-level attacks, since uploading an adversarial version of product images on a platform (e.g., eBay, Amazon, and Instagram) is more realistic than having access to the model in order to modify the visual feature used in both training/prediction phases. Furthermore, we focus our investigation on test-time/evasion attacks, assuming that, for an adversary, it can be easier and more efficient to change the product image on a platform to directly increase its predicted preference score and push it in high top- $K$  recommendation positions. In particular, considering the effects of items’ popularity on the recommendation quality [32, 23, 33], this work explored whether test time attacks might have different efficacy considering the items’ popularity.

## 3. Methodology

In this section, we first describe some useful notations, and then briefly review and present the formalization of the three adversarial attack strategies tested in the current work.

### 3.1. Preliminaries and Notations

**Recommendation Task.** Let  $\mathcal{U}$ ,  $\mathcal{I}$ , and  $\mathcal{S}$  be the set of users, items, and score-based preference feedback, where  $|\mathcal{U}|$ ,  $|\mathcal{I}|$ , and  $|\mathcal{S}|$  are the set sizes, respectively. Then, let  $s_{ui} \in \mathcal{S}$  be the preference of a user  $u \in \mathcal{U}$  on the item  $i \in \mathcal{I}$ , assuming that  $s_{ui} = 1$  when a user  $u$  has previously interacted with the item  $i$  (e.g. reviewed, purchased or clicked on the product). We define the recommendation task as the problem of producing a list of items that maximizes, for each user, a utility function. Moreover,  $\hat{s}_{ui}$  refers to the predicted preference score inferred from the RS trained on the set of user-item preference-feedback. A popular class of methods to learn unseen users' preferences is based on matrix factorization (MF) [34] techniques. The training of a MF-based recommender is aimed to learn an approximate version of the  $|\mathcal{U}| \times |\mathcal{I}|$  high-dimensional matrix of user-item preferences as the dot product of two low-rank matrices of latent factors. Each row of the first matrix is a *user latent* vector  $p_u \in \mathbb{P}^{|\mathcal{U}| \times h}$ , while each row of the second is the *item latent* vector  $q_i \in \mathbb{Q}^{|\mathcal{I}| \times h}$ , where  $h \ll |\mathcal{U}|, |\mathcal{I}|$ .

**Visual Feature Extraction in VRSs.** When it comes to visual recommendation, a common approach is to extract high-level visual features from (pretrained) CNNs (e.g., [4, 6, 28, 29]). We indicate with  $x_i$  the image/photo associated with the item  $i \in \mathcal{I}$ . While popular VRSs leverage either visual features extracted from pretrained CNNs or end-to-end trained approaches, we focus on the former class of visual recommenders, leaving the exploration of the latter as a future research direction. In this setting, given a set of data samples  $(x_i, y_i)$ , where  $x_i$  is the  $i$ -th image associated with the item  $i \in \mathcal{I}$  and  $y_i$  is the one-hot encoded representation of  $x_i$ 's image category, we indicate with  $F$  the DNN/CNN classifier pre-trained on all  $(x_i, y_i)$ . The network is trained such that the predicted probability vector of classes associated with an image ( $F(x_i) = \hat{y}_i$ ) is as much as close to the one-hot encoded vector of the ground truth-class  $y_i$ . Since the DNN is composed by  $L$ -layers, we indicate with  $F^{(l)}(x_i)$ ,  $0 \leq l \leq L - 1$ , the output of the  $l$ -th layer of  $F$  given the input  $x_i$ . The actual extraction takes place at one of the last layers of the network, i.e.,  $F^{(e)}(x_i)$ , where  $e$  refers to the extraction layer. In general, we define this layer output  $F^{(e)}(x_i) = \varphi_i$  as a mono-dimensional vector that will be the input to the VRS.

**A Popular Visual Recommender: VBPR.** To investigate the effects of items' popularity when affected by adversarial attacks, we considered the most popular baseline in the visually-aware recommendation task: Visual Bayesian Personalized Ranking from Implicit Feedback (VBPR) [4]. The model improves the MF preference predictor by adding a *visual* contribution to the traditional *collaborative* one. Given a user  $u$  and a non-interacted item  $i$ , the predicted preference score is  $\hat{s}_{ui} = p_u^T q_i + \theta_u^T \theta_i + \beta_{ui}$ , where  $\theta_u \in \Theta^{|\mathcal{U}| \times v}$  and  $\theta_i \in \Theta^{|\mathcal{I}| \times v}$  are the *visual* latent vectors of user  $u$  and item  $i$ , respectively ( $v \ll |\mathcal{U}|, |\mathcal{I}|$ ). The visual latent vector of item  $i$  is obtained as  $\theta_i = \mathbf{E}\varphi_i$ , where  $\varphi_i$  is the visual feature of image item  $i$  extracted from a pretrained convolutional neural network (i.e., AlexNet [8] as in the original work), while  $\mathbf{E}$  is a matrix to project the visual feature into the same space as of  $\theta_u$ . Furthermore,  $\beta_{ui}$  stands for the sum of the overall offset, the user, item, and global visual bias.

**The Failure Point of a Visual Recommender: the Visual Features.** All the state-of-the-art adversarial strategies [35, 19, 18, 17] alter the recommendation output ( $\hat{s}_{ui}^*$ ) by perturbing the products' images so that the newly-extracted visual feature  $\varphi_i^*$  leads to  $\hat{s}_{ui}^* \neq \hat{s}_{ui}$ . Then,  $\hat{s}_{ui}^* > \hat{s}_{ui}$  when the adversary wants to *push*/increase the score predicted on the target item  $i$  for the user  $u$ , while  $\hat{s}_{ui}^* < \hat{s}_{ui}$  holds in  *nuking* scenarios. To craft the adversarially perturbed

version of  $\varphi_i$ , the adversary can either have complete knowledge of the recommender model (i.e., parameters, output, and training data) or can be completely unaware of this information. In the former case, the adversary is generally recognized to work in **white-box** settings, while, in the latter case, she works in **black-box** ones.

**Adversarial Perturbation on Images.** We define an adversarial attack as the problem of finding the best value for a perturbation  $\delta_i$  such that the attacked image  $x_i^* = x_i + \delta_i$  must be visually similar to  $x_i$  according to a certain *distance* metric, e.g.,  $L_p$  norms, and  $x_i^*$  must stay within its original value range, i.e.,  $[0, 1]$  for 8-bit RGB images re-scaled by a factor 255. here, the intuition is that the visual feature extracted from the network ( $\varphi_i^* = F(x_i^*) \neq \varphi_i = F(x_i)$ ) will change the original behavior of the recommender towards the malicious goal. Independently on the adversary knowledge of the VRS, all the adversarial attack strategies defined in the literature and explored in this work learn the adversarial perturbation of a product image  $x_i$  (i.e.,  $\delta_i$ ) by backpropagating error information via  $F(\cdot)$ . Below, we define the three strategies to evaluate  $\delta_i$ : TAaMR, WB-SIGN, and INSA. Note that,  $x_i^*$  pixel values are clipped in the  $[0, 255]$  range of values at the end of the attack.

### 3.2. Black-Box: Targeted Adversarial Attack against Multimedia Recommenders (TAaMR)

The first adversarial attack strategy tested in this work is a test-time extension of the *Targeted Adversarial Attack against Multimedia Recommenders* attack (TAaMR) proposed by [16]. The strategy, originally proposed for poisoning the training set with altered items images, assumes to adversarially perturb the image such that the pretrained CNN used to extract the visual feature will misclassify the original images towards a different class. In particular, we use the Fast Gradient Sign Method (FGSM) [10] attack strategy—a baseline strategy in computer vision—that generates an adversarial version of the attacked image in only one step. Given a clean input image  $x_i$ , a target class  $p$ , a CNN  $F(\cdot)$  with parameters  $\theta$ , and a perturbation coefficient  $\epsilon$ , the targeted adversarial image  $x_i^*$  is:

$$x_i^* \leftarrow x_i - \epsilon \cdot \text{sign}(\nabla_{x_i} \mathcal{L}_F(\theta, x_i, p)) \quad (1)$$

where  $\nabla_{x_i} \mathcal{L}_F(\theta, x_i, y_i)$  is the loss function gradient of  $F(\cdot)$ , and  $\text{sign}(\cdot)$  is the sign function.

### 3.3. White-Box: Sign Method (WB-SIGN)

The second adversarial strategy is the *Sign-based White-Box Attack* (WB-SIGN) method that manipulates the product image computing the partial derivatives of the preference score function  $s(\cdot)$  with respect to the item image  $x_i$  used to predict the user-item preference score, and updates the pixels in that direction. Formally, we define the function of the sum of preference scores measured on all the items as:

$$\hat{s}_i(x_i) = \sum_{u \in \mathcal{U}} (\hat{s}_i(x_i)) = \sum_{u \in \mathcal{U}} (p_u^T q_i + \theta_u^T \mathbf{E}F(x_i) + \beta_{ui}) \quad \text{then} \quad x_i^* \leftarrow x_i - \epsilon \cdot \text{sign}\left(\frac{\partial \hat{s}_i(x_i)}{\partial x_i}\right) \quad (2)$$

### 3.4. White-Box: Insider Attack (INSA)

The last experimented attack is the *Insider Attack* (INSA) method proposed by Liu and Larson [18]. Similar to WB-SIGN, this method assumes that the adversary has full knowledge and the access to the parameters of the trained model, and uses it to modify the pixels from the product image to increase the preference scores inferred from the recommender on each target item. To this end, INSA is defined as follows:

$$x_i^* \leftarrow x_i - \frac{\partial \hat{s}_i(x_i)}{\partial x_i} \quad \text{such that} \quad \|\delta_i\| \leq \epsilon \quad \text{where} \quad \delta_i = \frac{\partial \hat{s}_i(x_i)}{\partial x_i} \quad (3)$$

Note that, similarly to WB-SIGN, INSA learns to build a perturbation that seeks to maximize all the users' scores predicted for each attacked item. However, differently from WB-SIGN, the INSA's perturbation is the gradient back-propagated through the recommender and the CNN, and not an  $\epsilon$ -bounded sign dependent perturbation.

## 4. Experiments

This section is devoted to presenting the setting we followed to run the experiments, and then discuss the obtained results.

### 4.1. Experimental Setup

**Datasets.** We perform the experiments on two recommendation datasets from [Amazon.com](https://www.amazon.com) containing customers' feedback and items' images. We use AMAZON BOYS & GIRLS and AMAZON MEN, including images in the fashion domain. The items set depend on the images still available on the e-commerce platform since they were not available in the released repository [24, 25]. We remove items/users with less than 5 interactions [5, 4]. Then, AMAZON BOYS & GIRLS has 1425, 5019, and 9213, while AMAZON MEN has 16278, 31750, and 113106, users, items, and feedback, respectively.

**Evaluation Metrics.** We evaluate the tested attacks according to the ability of the adversary to compromise the integrity of the recommendations, i.e., the efficacy in *increasing* the preference score predicted by the visual recommender and *pushing* the target items into the top- $K$  of each user's recommendation list. To measure the variation of the preference score, we define the Prediction Shift (PS) as follows:

$$\text{PS} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (\hat{s}_{ut}^* - \hat{s}_{ut}) \quad (4)$$

where  $\mathcal{T}$  is the set of target items whose images have been perturbed in an adversarial way. Additionally, to track the occurrences of an increase in the preference score, we measure the fraction of items for which we have measured a preference score improvement. We name this metric as Improvement Fraction (IF), and we formally define it as follows,

$$\text{IF} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} 1[\hat{s}_{ut}^* - \hat{s}_{ut}] \quad (5)$$



where  $IF < 0.5$  means that the number of times the target items have worsened their preference score is higher than the number of times it has been improved. While the previous metrics are related to preference score predictions, similar to [18, 17], we evaluate a ranking-wise metric that measures the average number of times a target item hits the top- $K$  recommendation lists. This metric, named Hit Ratio (HR@K), is defined as follows,

$$HR@K = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{U}|} \sum_{u \in |\mathcal{U}|} \text{hit}@K(t, u) \quad (6)$$

where  $\text{hit}@K(t, u)$  is 1 when the target item is in the top- $K$  list of the user  $u$ .

**Reproducibility.** We randomly initialize the model parameters with a Gaussian distribution with a mean of 0 and standard deviation of 0.01 and set the latent factor dimension to 128 as in [27]. We explore via grid-search: the learning rate in  $\{0.0001, 0.001, 0.01\}$  and the regularizers in  $\{0.00001, 0.001\}$ , whereas we fix the batch size to 256. We adopt early-stopping to avoid overfitting and choose the best model configuration for each algorithm according to the Recall@100 as in [27]. After having identified the best VBPR configuration on each dataset, we randomly sample 200 items from the catalog ( $|\mathcal{T}| = 200$ ). We attack each target item image and measure the correspondent adversarial score ( $\hat{s}_{ii}^*$ ) for each user. To study the effects of popularity on the attack efficacy, we split the target items into four groups based on the recorded feedback in the training set (i.e., Low Popular (LP), Mid-Low Popular (MLP), Mid-High Popular (MHP), and High Popular (HP)). For each attack, we vary the perturbation budget  $\epsilon \in \{4, 8, 16\}$ . For the black-box strategy (TAaMR), we select the target class as the most popular one (“running shoes”) for both datasets. All codes, datasets, and configuration files to run and evaluate the experiments are publicly available in the Elliot <sup>1</sup> reproducibility framework [17, 36].

## 4.2. Results and Discussion

In this section, we investigate the following research questions:

- RQ1:** What is the effect of items’ popularity on the efficacy of testing time adversarial attacks with respect to increasing the inferred preference scores? Is the behavior observed on a smaller perturbation budget ( $\epsilon = 4$ ) consistent with higher perturbation budgets ( $\epsilon \in \{8, 16\}$ )?
- RQ2:** While studying the adversary’s ability in pushing the target item into the top- $K$  recommendation list, what is the effect of items’ popularity? How much has  $\epsilon$  influenced ranking-wise performance?

## 4.3. Analysis of Attack Performance on Increasing Preference Scores (RQ1)

This paragraph analyses the Improvement Fraction (IF) and the Prediction Shift (PS) of the tested adversarial attacks against the AMAZON BOYS & GIRLS and AMAZON MEN datasets. Table 1 reports the measured performance on the four target items groups as defined in Section 4.1.

---

<sup>1</sup><https://github.com/sisinflab/elliott>

**Table 1**

Average Improvement Fraction and Prediction Shift across all popularity groups and perturbation budgets. We report the configuration files used to perform the experiments.

Attack	$\epsilon$	Improvement Fraction (IF)				Prediction Shift (PS)			
		LP	MLP	MHP	HP	LP	MLP	MHP	HP
AMAZON BOYS & GIRLS- [Elliot Configuration File: ws_attack_best_amazon_boys_girls.yml]									
TAaMR	4	0.44463	0.43154	0.41907	0.42219	-0.22579	-0.27033	-0.33856	-0.39528
	8	0.46338	0.44035	0.43328	0.42793	-0.14574	-0.24304	-0.30813	-0.36147
	16	0.49444	0.45799	0.47867	0.45004	-0.01750	-0.21294	-0.16771	-0.31063
WB-SIGN	4	0.90726	0.91332	0.90087	0.88025	2.02115	1.95690	1.93221	1.80204
	8	0.82153	0.82710	0.82013	0.78950	1.67599	1.61977	1.58222	1.44028
	16	0.70281	0.70766	0.71564	0.67601	1.15195	1.12273	1.11164	0.95283
INSA	4	0.98828	0.99128	0.98848	0.98801	0.89236	0.89898	0.80334	0.81315
	8	0.98831	0.99130	0.98841	0.98801	0.89228	0.89857	0.80303	0.81310
	16	0.98831	0.99130	0.98841	0.98801	0.89228	0.89844	0.80303	0.81310
AMAZON MEN- [Elliot Configuration File: ws_attack_best_amazon_men.yml]									
TAaMR	4	0.49325	0.49828	0.49779	0.42723	-0.03048	-0.04103	-0.00803	-0.40287
	8	0.45327	0.45121	0.46360	0.38513	-0.34333	-0.28482	-0.21996	-0.72266
	16	0.37303	0.38719	0.37847	0.33111	-1.03492	-0.85267	-0.88104	-1.32383
WB-SIGN	4	0.88149	0.85164	0.85136	0.82823	2.19162	2.02704	1.96209	1.91346
	8	0.74791	0.72156	0.72428	0.68396	1.45054	1.37203	1.35763	1.17692
	16	0.54332	0.55128	0.53101	0.50308	0.18964	0.31076	0.19959	0.06569
INSA	4	0.94537	0.91230	0.90757	0.90606	2.29845	2.09213	1.96360	2.05216
	8	0.93399	0.89413	0.89756	0.88974	2.21695	2.00421	1.91811	1.99631
	16	0.91968	0.87861	0.88227	0.87088	2.11271	1.91207	1.83971	1.91991

Before moving to the investigation of items' popularity effects of the adversarial attack efficacy, it is interesting to observe that the black-box strategy (TAaMR) is mostly ineffective in increasing the performance of the target items independently on the tested datasets. For instance, the most powerful attack with a perturbation budget equal to 16 has still negative PS values in both datasets (i.e., -0.01750 in <AMAZON BOYS & GIRLS, LP> and -0.00803 in <AMAZON MEN, MHP>) and IF values lower than 0.5 in all settings. We may explain the low efficacy of this attack strategy by stating that, differently from the TAaMR's proposal paper [37], we do not use the targeted adversarial strategy to poison the training procedure, but we use it in the testing phase when the model has already learned the users' preferences. However, we may derive two additional insights. The first one states that when increasing  $\epsilon$ , the attacks become more performant, and IF gets closer to 0.5. The second one states that the attack results reported for TAaMR have been more effective on low popular target items than the most popular ones. For instance, IF on **LP** items is higher than **HP** ones in all attack settings (e.g., 0.49444 > 0.45004 in AMAZON BOYS & GIRLS and 0.37303 > 0.33111 in AMAZON MEN with  $\epsilon = 16$ ).

After having discussed the black-box attack, we move the analysis to white-box ones. Starting from WB-SIGN, it can be noticed that the adversarial strategy is more effective against low popular items than high popular ones. For instance, PS with  $\epsilon = 4$  is 2.02115 on **LP** and 1.80204 on **HP** in AMAZON BOYS & GIRLS, while 2.19162 and 1.91346 on AMAZON MEN. The same result trends are confirmed for INSA, the secondly reported white-box attacks for both datasets, where, for example, 0.94537 > 0.91230 > 0.90757 > 0.90606 for the IF measured on AMAZON MEN



**Table 2**

HR@50 measured before the attack (*No Attacks*) and after the attack on all popularity groups. For each attack values, we report the percentage variation with respect to the not attacked version.

Attack	$\epsilon$	AMAZON BOYS & GIRLS			
		LP	MLP	MHP	HP
<i>No Attacks</i>		0.00669	0.01260	0.00762	0.01423
TAaMR	4	0.00392 (-70.97%)	0.00632 (-99.56%)	0.00519 (-46.76%)	0.00668 (-113.03%)
	8	0.00401 (-66.78%)	0.00627 (-100.89%)	0.00459 (-66.06%)	0.00622 (-128.89%)
	16	0.00536 (-24.87%)	0.00658 (-91.47%)	0.00504 (-51.25%)	0.00589 (-141.43%)
WB-SIGN	4	0.02429 (+72.44%)	0.0434 (+70.96%)	0.03168 (+75.94%)	0.04173 (+65.89%)
	8	0.01644 (+59.27%)	0.03148 (+59.96%)	0.02392 (+68.13%)	0.0296 (+51.92%)
	16	0.01241 (+46.04%)	0.02046 (+38.41%)	0.01697 (+55.09%)	0.02008 (+29.14%)
INSA	4	0.01321 (+49.31%)	0.02118 (+40.49%)	0.01533 (+50.27%)	0.0258 (+44.83%)
	8	0.01321 (+49.31%)	0.02118 (+40.49%)	0.01533 (+50.27%)	0.0258 (+44.83%)
	16	0.01321 (+49.31%)	0.02118 (+40.49%)	0.01533 (+50.27%)	0.0258 (+44.83%)
Attack	$\epsilon$	AMAZON MEN			
<i>No Attacks</i>		0.00044	0.00109	0.00091	0.00297
TAaMR	4	0.00089 (+50.28%)	0.00075 (-45.42%)	0.00105 (+13.64%)	0.0021 (-41.68%)
	8	0.00092 (+51.99%)	0.00103 (-5.83%)	0.00152 (+40.29%)	0.00224 (-32.51%)
	16	0.00082 (+46.04%)	0.00097 (-13.23%)	0.00162 (+43.82%)	0.00228 (-30.09%)
WB-SIGN	4	0.00299 (+85.19%)	0.00326 (+66.5%)	0.00527 (+82.74%)	0.01146 (+74.1%)
	8	0.00215 (+79.36%)	0.00279 (+60.76%)	0.00441 (+79.35%)	0.00741 (+59.92%)
	16	0.00143 (+69.04%)	0.00179 (+39.04%)	0.0029 (+68.63%)	0.0044 (+32.47%)
INSA	4	0.00285 (+84.45%)	0.00333 (+67.19%)	0.00478 (+80.97%)	0.01209 (+75.44%)
	8	0.00287 (+84.57%)	0.00347 (+68.46%)	0.00507 (+82.05%)	0.01206 (+75.38%)
	16	0.00281 (+84.22%)	0.00335 (+67.32%)	0.0047 (+80.63%)	0.01195 (+75.15%)

with  $\epsilon = 4$ . The same trends on both WB-attacks and datasets are observed when varying  $\epsilon$ .

These empirical observations confirm that *items' popularity has influenced the efficacy of adversarial attack strategies, where the least popular target items are subject to an increment of the preference scores much bigger than those calculated on the most popular ones. In addition, the tendency mentioned above is consistent when varying the perturbation budget from small values (i.e., 4) to larger ones (i.e., a maximum value of 16).*

#### 4.4. Analysis of Attack Performance on Top-K Recommendation Lists (RQ2)

Table 2 reports the HR@50 values measured on top-50 recommendation lists before and after the execution of adversarial attacks. This paragraph seeks to verify if the higher attack efficacy on low popular items measured from a preference score point of view is consistent when analyzing the top- $K$  recommendation lists. Differently from the previous analysis, we should point out that most popular items may have HR@50 values (before the attack) higher than the ones of low popular items due to the well-known popularity bias issues [20, 21, 22, 23]. Indeed, it can be observed that the HR@50 of **HP** items is more than two times higher than the one measured on **LP** (i.e.,  $0.01423 > 0.0069$ ) in AMAZON BOYS & GIRLS, and even more than six times higher in AMAZON MEN (i.e.,  $0.00297 > 0.00044$ ). For this reason, we report the HR@50 variation after the attack in Table 2.

Analyzing the variations measured under black-box attack settings, it can be noted that,

consistently with the findings measured in Section 4.3, low popular target items have been more affected by attacks than most popular ones. Indeed, despite the negative variations of HR@50 measured on **LP** items in AMAZON BOYS & GIRLS (i.e., -70.97%), the ones measured on **HP** are even more negative (i.e., -113.03). The same trend is also confirmed in the AMAZON MEN dataset independently of the perturbation budget. Extending the analysis to the white-box adversarial strategies, and considering that PS is always greater than 1 and IF is greater than 0.5, we should expect that the percentage variations measured on **LP** and **MLP** items should be higher than the ones on **MHP** and **HP**. Results in Table 2 confirm that both WB-SIGN and INSA are more effective on **LP** items. For instance, HR@50 increases by +85.19% on **LP** and +74.10% on **HP**, when WB-SIGN with  $\epsilon = 4$  is performed on AMAZON MEN.

Results on the top- $K$  recommendation performance additionally confirm that *items' popularity affects the efficacy of attacks by making the least popular target items easier to push into higher positions than the ones already in high positions.*

## 5. Conclusion

We examined if test-time adversarial attacks against VRSs have a distinct impact on items based on their popularity. To this end, we tested one black-box (i.e., TAaMR) and two state-of-the-art white-box (i.e., WB-SIGN and INSA) single-step adversarial attacks by varying three levels of perturbation budget (i.e.,  $\epsilon \in \{4, 8, 16\}$ ) to alter the recommendations generated by VBPR, a baseline model for visual recommendation. Indeed, after VBPR's training on two datasets (i.e., AMAZON BOYS & GIRLS and AMAZON MEN), we randomly extracted 200 target items from each catalog. Then, we divided them into four groups based on the number of ratings registered in the training set, and performed two analyses: one on the preference score and the other on the effects on top- $K$  lists. From the former, we found that items' popularity influences the attacks' efficacy, which is much more effective on the least popular than high popular items in incrementing the preference scores consistently at varying of  $\epsilon$ . From the latter, we verified that this trend is also confirmed when looking at top- $K$  recommendation lists, with the least popular target items getting the highest pushing in ranking positions. These results open exciting challenges for developing adversarial defenses strategies, as the least popular items can be highly subjected to adversarial attacks. We propose extending the study on iterative adversarial attacks to understand if the previously identified trends are consistent with more robust strategies for future extension. Finally, we plan to extend this investigation line to examine the potential effects of users' activeness (e.g., number of released ratings) on the attack efficacy for providing more insights into planning more powerful defense strategies and to study of the verified effects with human evaluation.

## Acknowledgments

We acknowledge support of PON ARS01\_00876 BIO-D, Casa delle Tecnologie Emergenti della Città di Matera, PON ARS01\_00821 FLET4.0, PIA Servizi Locali 2.0, H2020 Passapartout - Grant n. 101016956, and PIA ERP4.0.

## References

- [1] Y. Hu, X. Yi, L. S. Davis, Collaborative fashion recommendation: A functional tensor factorization approach, in: *ACM Multimedia*, ACM, 2015, pp. 129–138.
- [2] D. Elsweiler, C. Trattner, M. Harvey, Exploiting food choice biases for healthier recipe recommendation, in: *SIGIR*, ACM, 2017, pp. 575–584.
- [3] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, H. Liu, What your images reveal: Exploiting visual contents for point-of-interest recommendation, in: *WWW*, ACM, 2017, pp. 391–400.
- [4] R. He, J. J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback, in: *AAAI*, AAAI Press, 2016, pp. 144–150.
- [5] R. He, J. J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: *WWW 2016*, 2016.
- [6] Q. Liu, S. Wu, L. Wang, Deepstyle: Learning user preferences for visual recommendation, in: *SIGIR*, ACM, 2017, pp. 841–844.
- [7] L. Meng, F. Feng, X. He, X. Gao, T. Chua, Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation, in: *ACM Multimedia*, ACM, 2020, pp. 3460–3468.
- [8] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *NeurIPS 2012*, 2012.
- [9] Y. Deldjoo, T. D. Noia, F. A. Merra, A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks, *ACM Computing Surveys (CSUR)* (2021).
- [10] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: *ICLR (Poster)*, 2015.
- [11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: *ICLR 2018*, 2018.
- [12] N. Carlini, D. A. Wagner, Towards evaluating the robustness of neural networks, in: *SP 2017*, 2017.
- [13] X. He, Z. He, X. Du, T. Chua, Adversarial personalized ranking for recommendation, in: *SIGIR*, ACM, 2018, pp. 355–364.
- [14] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, T. Chua, Adversarial training towards robust multimedia recommender system, *IEEE Trans. Knowl. Data Eng.* 32 (2020) 855–867.
- [15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *CVPR*, IEEE Computer Society, 2016, pp. 770–778.
- [16] T. Di Noia, D. Malitesta, F. A. Merra, Taamr: Targeted adversarial attack against multimedia recommender systems, in: *DSN–DSML 2020*, 2020.
- [17] V. W. Anelli, Y. Deldjoo, T. Di Noia, D. Malitesta, F. A. Merra, A study of defensive methods to protect visual recommendation against adversarial manipulation of images, in: *SIGIR*, ACM, 2021, pp. 1094–1103.
- [18] Z. Liu, M. A. Larson, Adversarial item promotion: Vulnerabilities at the core of top-n recommenders that use images to address cold start, in: *WWW, ACM / IW3C2*, 2021, pp. 3590–3602.
- [19] R. Cohen, O. S. Shalom, D. Jannach, A. Amir, A black-box attack model for visually-aware

- recommender systems, in: WSDM, ACM, 2021, pp. 94–102.
- [20] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, *User Model. User Adapt. Interact.* 25 (2015) 427–491.
  - [21] H. Abdollahpouri, R. Burke, B. Mobasher, Controlling popularity bias in learning-to-rank recommendation, in: *RecSys*, ACM, 2017, pp. 42–46.
  - [22] Z. Zhu, J. Wang, J. Caverlee, Measuring and mitigating item under-recommendation bias in personalized ranking systems, in: *SIGIR*, ACM, 2020, pp. 449–458.
  - [23] L. Boratto, G. Fenu, M. Marras, Connecting user and item perspectives in popularity debiasing for collaborative recommendation, *Inf. Process. Manag.* 58 (2021) 102387.
  - [24] R. He, C. Packer, J. J. McAuley, Learning compatibility across categories for heterogeneous item recommendation, in: *ICDM*, IEEE Computer Society, 2016, pp. 937–942.
  - [25] J. J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: *SIGIR 2015*, 2015.
  - [26] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. Di Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: *SIGIR*, ACM, 2021, pp. 2405–2414.
  - [27] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, T. Chua, Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention, in: *SIGIR*, ACM, 2017.
  - [28] W. Niu, J. Caverlee, H. Lu, Neural personalized ranking for image recommendation, in: *WSDM 2018*, 2018.
  - [29] W. Kang, C. Fang, Z. Wang, J. J. McAuley, Visually-aware fashion recommendation and design with generative image models, in: *ICDM*, IEEE Computer Society, 2017, pp. 207–216.
  - [30] R. Yin, K. Li, J. Lu, G. Zhang, Enhancing fashion recommendation with visual compatibility relationship, in: *WWW 2019*, 2019.
  - [31] A. Paul, Z. Wu, K. Liu, S. Gong, Robust multi-objective visual bayesian personalized ranking for multimedia recommendation, *Applied Intelligence* (2021) 1–12.
  - [32] R. Cañamares, P. Castells, Should I follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems, in: *SIGIR*, ACM, 2018, pp. 415–424.
  - [33] E. Mena-Maldonado, R. Cañamares, P. Castells, Y. Ren, M. Sanderson, Popularity bias in false-positive metrics for recommender systems evaluation, *ACM Trans. Inf. Syst.* 39 (2021) 36:1–36:43.
  - [34] Y. Koren, R. M. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, *Computer* 42 (2009) 30–37.
  - [35] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, T. Chua, Adversarial training towards robust multimedia recommender system, *IEEE Trans. Knowl. Data Eng.* 32 (2020) 855–867.
  - [36] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. Di Noia, V-elliot: Design, evaluate and tune visual recommender systems, in: *RecSys*, ACM, 2021.
  - [37] T. Di Noia, D. Malitesta, F. A. Merra, Taamr: Targeted adversarial attack against multimedia recommender systems, in: *DSN Workshops*, IEEE, 2020, pp. 1–8.