TAaMR: Targeted Adversarial Attack against Multimedia Recommender Systems

Tommaso Di Noia
Politecnico di Bari
tommaso.dinoia@poliba.it

Daniele Malitesta
Politecnico di Bari
daniele.malitesta@poliba.it

Felice Antonio Merra
Politecnico di Bari
felice.merra@poliba.it

***Abstract*—Deep learning classifiers are hugely vulnerable to adversarial examples, and their existence raised cybersecurity concerns in many tasks with an emphasis on malware detection, computer vision, and speech recognition. While there is a considerable effort to investigate attacks and defense strategies in these tasks, only limited work explores the influence of targeted attacks on input data (e.g., images, textual descriptions, audio) used in multimedia recommender systems (MR). In this work, we examine the consequences of applying targeted adversarial attacks against the product images of a visual-based MR. We propose a novel adversarial attack approach, called Target Adversarial Attack against Multimedia Recommender Systems (*TAaMR*), to investigate the modification of MR behavior when the images of a category of low recommended products (e.g., socks) are perturbed to misclassify the deep neural classifier towards the class of more recommended products (e.g., running shoes) with human-level slight images alterations. We explore the *TAaMR* approach studying the effect of two targeted adversarial attacks (i.e., FGSM and PGD) against input pictures of two state-of-the-art MR (i.e., VBPR and AMR). Extensive experiments on two real-world recommender fashion datasets confirmed the effectiveness of *TAaMR* in terms of recommendation lists changing while keeping the original human judgment on the perturbed images.**

***Index Terms*—Adversarial Machine Learning, Recommender Systems**

I. Introduction

Deep Neural Networks (DNN) serve as core components of many real-world systems for performing different AI tasks such as image classification [1], object detection [2], speech recognition [3], and malware detection [4]. However, recent studies have demonstrated that a malicious user, the adversary, can modify the classification behavior of a trained deep neural classifier by attaching human-imperceptible adversarial noise on inputs at prediction time [5]. A famous example in the computer vision domain is the misclassification of a slightly mutated STOP traffic signal into another one by a DNN classifier installed in a self-driving car system [6]. Moreover, recent researches have proved that adversaries might have the capabilities to generate adversarial examples such that they are misclassified into a chosen target class, performing the so-called targeted adversarial attacks [7], [8].

The power of DNN in providing latent representations (features) of input data in a supervised and unsupervised way has been recently exploited in the application domain of recommender systems (RS). They act as a primary component in several real-world online product retailers (e.g., Amazon [9]) or multimedia content providers (e.g., Netflix [10]) by furnishing users with personalized recommendations to simplify the identification of the product best suiting their preferences in catalogs with millions of potential alternatives. Besides, the availability of several types of multimedia content for products/services (e.g., images [11], [12], video [13], soundtracks [14]) supports RS to have better-personalized recommendations.

Recommendation engines are prone to performance alteration by malicious users that might be able to poison the training data with hand-engineered, and machine-learning optimized, fake user profiles (*shilling profiles*). This attack scenario —comprehensively studied in the literature [15]–[17]— is different from the adversarial machine learning one since adversarial perturbations are evaluated in an optimized way to be human-imperceptible. Recently, adversarial attacks (and defense) have been studied in RS [18] with a primary concern on the evaluation of adversarial perturbations applied to recommender model embeddings [19]. Indeed, this approach is also used in [20], the first work to explore the effectiveness of adversarial attacks against MR (i.e., a visual-based recommender model). Differently from our work, the authors of [20] investigated the performance worsening with **untargeted** perturbation on input images.

In this work, we show how the performance of MR is effectively modified by an adversary that might insert **targeted** adversarial perturbed images in a visual-based recommender system (e.g., VBPR [12]). The proposed attack approach, named Targeted Adversarial Attack against Multimedia Recommender Systems (*TAaMR*), explores attack situations where the adversary's goal is to perturb images of a low recommended category of products (e.g., the 20th most recommended) to be misclassified by the deep classifier towards a target more recommended category (e.g., the 1st/2nd).

In brief, this work aims at addressing the following research questions:

RQ1 Can targeted adversarial attacks against images of a low recommended category of products be exploited to modify the recommendation lists of multimedia recommender systems in terms of probability of being more recommended?

RQ2 What are the effects of adversarial perturbations against these attacked images for human-perceptions?

To answer the previous questions, we have performed an extended experimental evaluation to investigate the effects of *TAaMR* on two real-world visual-based fashion datasets (`Amazon Men` and `Amazon Women`) where visual features are predominant. The code is available at `https://github.com/sisinflab/TAaMR`.

II. Problem Definition

The discovery of a utility function able to predict how much a user (e.g., an e-commerce client) will like an unknown-item (e.g., a product for sale) is the central recommendation problem task.

Definition 1 (Recommendation Problem). *Let $\mathcal{U}$ and $\mathcal{I}$ denote a set of users and items in a RS, respectively, and $c : \mathcal{U} \times \mathcal{I} \rightarrow \mathbb{R}$ be a utility function. The* ***Recommendation Problem*** *is defined as*

$$\forall u \in \mathcal{U}, \hat{i}_u = \arg\max_{i \in \mathcal{I}} c(u, i) \qquad (1)$$

where $\hat{i}_u$ is the recommended item to the user u.

We further define $S \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ as the user-item feedback matrix (UFM), where each entity $s_{ui} \in \mathbb{R}$ is a 0/1-valued feedback (e.g., review, rating) that represents an historical interactions for user $u \in \mathcal{U}$ to item $i \in \mathcal{I}$.

Definition 2 (Learned Image Feature). *Let $\mathcal{X}$ be a set of images and F a trained DNN model (e.g., CNN). Then, let L be the number of layers of F, let f^l be the output of the l-th layer of F with $l = \{0, 1, ..., L-1\}$ and $F(\cdot) = f^{L-1}$. Given an image $x \in \mathcal{X}$, the* ***Learned Image Feature*** *of x, extracted at layer l is defined as $f^l(x)$.*

Given an item i associated to an image x_i, we represent its features f_i by extracting the output f^e of a layer e. In this work, we select e as one of the layers placed immediately after the convolutional part since all the stacked convolutions have extracted high-level features from images that are used in a multimedia *Recommendation Problem*.

Adversarial attacks can be untargeted or targeted. An untargeted adversarial strategy is only interested in a perturbation of the input x in x^* such that the predicted class $F(x^*)$ for x^* is different from the original one $F(x)$.

Definition 3 (Untargeted Adversarial Attack). *An* ***Adversarial Attack*** *against a DNN (e.g., CNN) is the process of finding an* ***adversarial example*** *x^*, solving the following constrained optimization problem:*

$$\begin{aligned} &\min_{\mathrm{d} \leq \epsilon} d(x, x^*) \\ &\textit{such that } F(x) \neq F(x^*) \end{aligned} \qquad (2)$$

where x is a clean sample for F, $\mathrm{d}(\cdot)$ is a distance metric, and $\epsilon > 0$ is the perturbation budget.

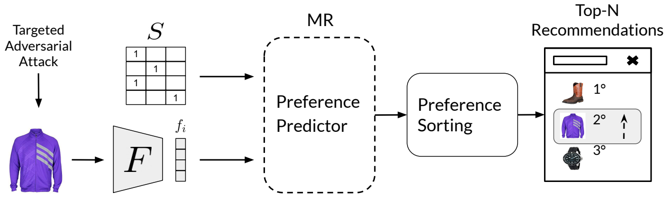


Fig. 1: Overview of *TAaMR*.

Untargeted adversarial attacks construct adversarial samples (i.e., images) *maximizing* the classification cost function related to the original (source) class. On the other side, a targeted strategy is interested in changing the original class of the perturbed example to a specific new one t.

Definition 4 (Targeted Adversarial Attack). *Let $\mathcal{C}$ be a set of classes for a classifier F. Let $c \in \mathcal{C}$ be the source class such that $F(x) = c$, and $t \in \mathcal{C}$ be a target class with $t \neq c$. A* ***Targeted Adversarial Attack*** *finds the adversarial examples x^* as following:*

$$\begin{aligned} &\min_{\mathrm{d} \leq \epsilon} d(x, x^*) \\ &\textit{such that } F(x^*) = t \end{aligned} \qquad (3)$$

Targeted adversarial attacks generate adversarial images to *minimize* the classification loss function with respect to the target class. Based on the previous problem definitions, we propose a novel attack strategy against a visual-based recommender system, called *Targeted Adversarial Attack against Multimedia Recommender Systems* (*TAaMR*).

III. Approach

In the description of our approach on perturbing input images over multimedia recommenders (MR) with targeted adversarial attacks, we first introduce the core multimedia recommender model, and then we define the adversarial threat model. Finally, we discuss a novel metric proposed to evaluate the impact of the attack on the recommendation lists. In Fig. 1 we visually represent the approach.

A. Multimedia Recommender

The core component of *TAaMR* is the multimedia recommender. A multimedia recommender model solves the Recommendation Problem estimating the user preferences of unknown items/products combining pure collaborative filtering information (i.e., users' interactions) with multimedia-based features. As shown in Fig. 1, the proposed approach investigates the class of MR that performs the preference prediction task integrating features extracted from deep neural models (e.g., CNN in the case of visual-based MR).

The deep feature extractor component (F) is the vulnerability point exploited by an adversary to have a direct influence on the preference predictor component of the MR. The basic intuition of *TAaMR* is that an adversary could tamper the Recommendation Problem by abusing the so far demonstrated deficiency of several DNN to targeted adversarial examples. Indeed, *TAaMR* simulates targeted attacks on images in low recommended categories towards highly popular target categories.

B. Adversary Threat Model

Before diving into the examination of the consequences of the source-target-misclassification attack [7] (or targeted attack [8]) on a MR, we outline the adversary threat model based on the guidelines proposed by Carlini et al. [21]. The adversary's assumptions are:

- **adversary goal:** The adversary is interested in misclassifying images of a low suggested category of products from their source class (e.g., socks) towards a target one (e.g., t-shirts).
- **adversary knowledge:** We assume a white-box knowledge setting since the adversary holds a full knowledge of the feature extraction model parameters used to estimate the targeted perturbation. Additionally, the adversary has complete access to the MR input image features altered due to the performed attack. Furthermore, she can extract all the recommendation lists used to identify the source-target classes of *TAaMR*.
- **adversary capability:** We restrict the adversary capability to make l_∞-norm constrained perturbations.

C. Impact on Recommendations

Prior research in adversarial machine learning focused on the evaluation of adversarial attacks in discrediting the classification accuracy of 'victim' classifiers [22] (e.g., CNN), and the validation of defense strategies [21], [22]. To the best of our knowledge, there are not metrics to examine the impact of targeted adversarial attacks against MR. As evidence, the closest line of research to our work [20], evaluates the effects of untargeted attacks as the reduction in recommendation accuracy metrics [23], [24].

In this work, we aim to fill the mentioned gap by proposing a Hit Ratio-based metrics, named Category Hit Ratio ($CHR@N$), to study the fraction of the category of attacked items — whose images have been adversarially perturbed — in the top-N recommendation lists.

Definition 5 (Category Hit Ratio). *Let $c \in \mathcal{C}$ be a category (class), and $I_c = \{i \in I | F(x_i) = c\}$. The Category Hit Ratio@N is defined as*

$$CHR@N(I_c, U) = \frac{1}{N \cdot |U|} \sum_{u \in U} \sum_{i \in I_c \setminus I_u^+} hit(i, u) \tag{4}$$

where $hit(i, u)$ is a 0/1-valued function that is 1 when the item i, classified within the c class, is inside the top-N recommendation list of the user u, otherwise it is 0.

IV. Experimental Evaluation

We evaluated *TAaMR* on two real-world datasets in the recommendation domain. Firstly, we present the experimental setup; then we discuss the experimental results. We have carried out an extensive set of experiments to answer the research questions raised in Section I, i.e., whether it is possible to (i) modify the MR recommendation list by perturbing input images with targeted attacks, (ii) evaluate the visual appearance of perturbed product images.

TABLE I: Dataset statistics. $|U|$, $|I|$, and $|S|$ represent the number of users, items and feedback respectively.

Dataset	$\|U\|$	$\|I\|$	$\|S\|$
`Amazon Men`	$26,155$	$82,630$	$193,365$
`Amazon Women`	$18,514$	$76,889$	$137,929$

A. Evaluation Settings

1) Dataset: We executed experiments on two popular datasets extracted from *Amazon.com* [11], [25]. We considered the "Clothing, Shoes and Jewelry" category of men and women, named `Amazon Men` and `Amazon Women`, since several works demonstrated the significative impact of visual-features on users' choices in the fashion domain [12], [26]. Table I shows the dataset statistics as a result of preprocessing steps applied to each dataset. As a first step, we have downloaded all the available images from *Amazon.com* based on the URL published within the available metadata (`http://jmcauley.ucsd.edu/data/amazon/`). Then, we have converted any users' rating into a 0/1-valued interaction, and we have considered all the users with at least five interactions ($|I_u^+| \geq 5$) to discard cold-users. Furthermore, we have produced a smaller version of `Amazon Women` to make the number of product images comparable with `Amazon Men`.

2) Adversarial Attacks: We took into account two state-of-the-art adversarial attacks, i.e., Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).

Fast Gradient Sign Method (FGSM) [27] focuses on the speed of the adversarial example generation. As a matter of fact, it generates an adversarial version of the attacked image in only one step. Given a clean input image x, a target class t, a model F with parameters θ, and a perturbation coefficient ϵ, the targeted adversarial image x^* is then

$$x^* \leftarrow x - \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}_F(\theta, x, t)) \tag{5}$$

where $\nabla_x \mathcal{L}_F(\theta, x, y)$ is the gradient of the F loss function, and $\mathrm{sign}(\cdot)$ is the sign function.

Projected Gradient Descent (PGD) [28] iteratively applies FGSM, with a budget perturbation α (i.e., the step size) smaller than ϵ. The attack algorithm works similarly to FGSM, but after each completed perturbation step, the temporary attacked image is clipped to remains in a ϵ-neighborhood of the clean image x. The described approach is an extended and more effective version [29] of Basic Iterative Method (BIM) proposed by Kurakin et al. [30]. Indeed, PGD differs from BIM in the fact that PGD starts from a uniform random noise as the initial perturbation on the clean image x. The implemented version executes 10 iterations. For the implementation of both the algorithms, we adopted the Python library CleverHans [31]. Both the attacks have been executed in their targeted version.

3) Recommender Models: We studied *TAaMR* effectiveness on the following MR: Visual Bayesian Personalized Ranking (VBPR), and Adversarial Multimedia Recommendation (AMR).

Visual Bayesian Personalized Ranking (VBPR) [12] is a state-of-the-art multimedia recommender model designed to integrate visual features (learned via CNNs) into a latent factor model (BPR-MF [32]). The fundamental idea of VBPR is that a user might be influenced by the visual appearance of product images (e.g., a T-shirt picture on *Amazon.com*).

The preference predictor of VBPR (see Fig 1) is built on top of a matrix factorization (MF) model [33]. Given the $\mathcal{D}$-dimensional visual feature f_i of an item i, $\mathbf{E}$ representing a $\mathcal{D} \times \mathcal{A}$ matrix to transform f_i into a smaller $\mathcal{A}$-dimensional latent representation, and u as a user who did not interact with i, then the preference score ($\hat{s}_{ui}$) is calculated as:

$$\hat{s}_{ui} = b_{ui} + p_u^T q_i + \alpha_u^T (\mathbf{E} f_i) + \beta^T f_i \tag{6}$$

where p_u and q_i are $\mathcal{K}$-dimensional ($\mathcal{K} << |U|, |I|$) latent representations of user u and item i, b_{ui} is the sum of the global offset, the user, and item biases components, and β is a parameter to represent the overall effect of visual features on users preferences.

VBPR estimates model parameters θ by minimizing a pairwise ranking loss [32]. The basic intuition is that, given a triplet (u, i, j) of a user $u \in U$, an interacted item $i \in I_u^+$ and a not-interacted item $j \in I_u^-$, where I_u^+ and I_u^- are, respectively, the set of interacted and not-interacted items by the user u; then the preference score $\hat{s}_{ui}$ should be higher than $\hat{s}_{uj}$. Let $T = \{(u, i, j) | u \in \mathcal{U}, i \in \mathcal{I}_u^+, j \in \mathcal{I}_u^-\}$ be the set of triplets, then the VBPR optimization problem consists in minimizing the following objective function:

$$\mathcal{L}_{VBPR} = \sum_{(u,i,j) \in T} - \ln \sigma(\hat{s}_{ui} - \hat{s}_{uj}) + \lambda \|\theta\|_2^2 \tag{7}$$

where λ is the regularization coefficient of the L_2-norm of model parameters, and $\sigma(\cdot)$ is the sigmoid function.

Adversarial Multimedia Recommendation (AMR) [20] integrates VBPR with the adversarial training procedure for RS [19] to make the model more robust to adversarial perturbation (Δ_i) applied on the i-th image feature f_i. The preference prediction function is determined as:

$$\hat{s}_{ui}^{adv} = b_{ui} + p_u^T q_i + \alpha_u^T (\mathbf{E}(f_i + \Delta_i)) + \beta^T (f_i + \Delta_i) \tag{8}$$

where $\Delta_i \in \mathbb{R}^{1 \times \mathcal{D}}$ is the optimal adversarial perturbation that maximizes (7). Based on the FGSM-like adversarial attack proposed in [19], the matrix of adversarial perturbations on image features $\Delta_{adv} \in \mathbb{R}^{|I| \times \mathcal{D}}$ is evaluated as:

$$\Delta_{adv} = \eta \frac{\Pi}{\|\Pi\|} \quad \text{where} \quad \Pi = \frac{\partial \mathcal{L}_{VBPR}(T | \hat{\theta} + \Delta)}{\partial \Delta} \tag{9}$$

where $\hat{\theta}$ represents the fixed model parameters, η is the coefficient to control the magnitude of the feature perturbation, and Δ is the zeros-initialized perturbation matrix.

To reduce the impact of the introduced perturbations, AMR learns parameters θ minimizing the objective function:

$$\begin{aligned} \mathcal{L}_{AMR} &= \sum_{(u,i,j) \in T} - \ln \sigma(\hat{s}_{ui} - \hat{s}_{uj}) - \gamma \ln \sigma(\hat{s}_{ui}^{adv} - \hat{s}_{uj}^{adv}) \\ &= \mathcal{L}_{VBPR}(T|\theta) + \gamma \underbrace{\mathcal{L}_{VBPR}(T|\theta + \Delta_{adv})}_{\text{adversarial regularizer}} \end{aligned} \tag{10}$$

where γ is a weight coefficient to control the impact of the *adversarial regularizer*.

We have trained VBPR for 4000 epochs storing the model parameters at 2000-th epoch,i.e., the point where AMR starts the further 2000-epochs of adversarial training based on 10. VBPR and AMR hyper-parameters are set based on the configuration proposed in [20]; in particular, the parameters of the *adversarial regularizer* are set to $\gamma = 0.1$ and $\eta = 1$.

4) Visual Evaluation Metrics: We need to evaluate the visual distortion between the original and the attacked images since they are presented to a real online customer. There exist several quality metrics to measure the amount of distortion between two images (i.e., x and x^*), categorized as *subjective* and *objective* [34]. The former involves the quality evaluation of actual human users, but they are not always easy to collect. Conversely, the latter aims at mathematically mimicking a human evaluation. In this work, we studied the following objective metrics: Peak Signal-To-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and a Perceptual Similarity Metric (PSM).

Peak Signal-To-Noise Ratio (PSNR) [35] is a more easily interpretable, logarithmic version of the Mean Squared Error (MSE) and is defined as:

$$\text{PSNR}(x, x^*) = 10 \log_{10} \left(\frac{P^2}{\text{MSE}(x, x^*)} \right) \tag{11}$$

where P is the maximum pixel value (e.g., $P = 255$ for 8-bit images). The higher the PSNR value, the lower the distortion between x and x^*. Typically, it ranges between 20 and 50 dB.

Structural Similarity Index (SSIM) [36] is an objective metric based on the assumption that humans are sensitive to the *structure* of the image, an aspect that PSNR (as well as MSE) is not always able to capture [36]. The approach calculates local SSIM indexes over smaller windows of the original images, and finally takes the average. Given two corresponding image windows w and w^* extracted respectively from x and x^*, the SSIM between w and w^* is:

$$\text{SSIM}(w, w^*) = \frac{(2\mu_w \mu_{w^*} + k_1)(2\sigma_{ww^*} + k_2)}{(\mu_w^2 + \mu_{w^*}^2 + k_1)(\sigma_w^2 + \sigma_{w^*}^2 + k_2)} \tag{12}$$

where μ_v and σ_v are the mean and the standard deviation of the v-th window, σ_{ww^*} is the cross-correlation between w and w^*, and k_1, k_2 are constants to avoid a zero denominator. SSIM index ranges from in $-1, \ldots, 1$, with 1 denoting the perfect structural similarity.

Perceptual Similarity Metric (PSM) is a perceptual loss-based metric to evaluate the similarity of two images by leveraging on their *high-level* features extracted from pre-trained CNN. The idea is to access the *semantic content* of the images rather than their raw pixel-values [37]. We use the *feature reconstruction loss* from [37], but (i) we use a different pre-trained CNN and (ii) we exploit the same output layer (e) selected for the recommender. PSM is defined as:

$$\text{PSM}(x, x^*) = \frac{1}{H_e W_e C_e} \left\| f^e(x) - f^e(x^*) \right\|_2^2 \tag{13}$$

where H_e, W_e and C_e are the height, width and channel dimensions of f^e.

TABLE II: *TAaMR* experimental results ($CHR@100$).

Dataset	MRS	Attack	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Amazon Men	VBPR		Sock(2.122)→Running Shoes(7.888)				Sock(2.122)→Analog Clock(4.760)			
		FGSM	2.131	2.595	2.994	3.500	1.617	1.840	2.038	2.120
		PGD	3.654	5.562	**6.402**	5.931	1.663	2.072	3.168	**4.348**
	AMR		Sock(2.088)→Running Shoes(8.339)				Sock(2.088)→Jersey, T-shirt(10.745)			
		FGSM	2.015	**2.548**	1.910	1.964	1.502	1.717	1.475	1.755
		PGD	2.050	2.219	2.520	2.210	1.871	1.710	1.835	**1.980**
Amazon Women	VBPR		Maillot(1.546)→Brassiere(6.224)				Maillot(1.546)→Chain(4.374)			
		FGSM	1.500	1.529	1.567	1.538	1.092	1.193	1.459	1.279
		PGD	2.470	3.153	3.965	**4.421**	1.014	1.432	3.609	**4.992**
	AMR		Maillot(1.537)→Brassiere(10.816)				Maillot(1.537)→Chain(2.728)			
		FGSM	**1.990**	1.715	1.843	1.674	2.093	1.720	1.916	1.605
		PGD	1.136	1.168	0.827	1.006	2.228	**2.516**	1.863	1.912

TABLE III: Attacks success probability.

Dataset	Origin→Target	Attack	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Amazon Men	Sock→Running Shoes	FSGM	9.32%	17.02%	22.14%	21.68%
		PGD	68.69%	**98.37%**	**99.92%**	**99.84%**
	Sock→Analog Clock	FSGM	0.16%	0.31%	0.39%	0.23%
		PGD	30.77%	87.10%	**99.46%**	**100.00%**
	Sock→Jersey, T-shirt	FGSM	8.24%	17.17%	26.50%	15.54%
		PGD	67.29%	**98.83%**	**100.00%**	**100.00%**
Amazon Women	Maillot→Brassiere	FGSM	45.51%	51.48%	52.30%	56.46%
		PGD	85.32%	**99.40%**	**99.95%**	**100.00%**
	Maillot→Chain	FGSM	0.38%	1.31%	1.92%	2.68%
		PGD	17.20%	**90.53%**	**99.95%**	**99.95%**

5) Experimental Protocol: For each dataset, the evaluation protocol started with the calculation of the $CHR@100$ on the recommender models (i.e., VBPR, and AMR) trained on *clean* images. Based on the initial $CHR@100$, we selected two attack scenarios. The first simulates attacks on source-target categories that are semantically similar (e.g., Sock→Running Shoes), while the second assumes semantically different classes (e.g., Sock→Analog Clock). We have selected *Sock → Jersey, T-Shirt* for AMR on Amazon Men since *Analog Clock* category is not as highly recommended as in VBPR model. After setting the source-target categories, we implemented the adversarial attacks (i.e., FGSM and PGD) described in Section IV-A2. To achieve a fair comparison between both attacks, we set the same magnitude for the coefficient ϵ (i.e., $\epsilon = \{2, 4, 8, 16\}$, with ϵ normalized to a 0/1 scale); while the remaining parameters have been set with the default values in CleverHans [31]. The experimented CNN is ResNet50 [38]. We set e (the layer to extract f_i) as the output of the global average pooling right after the convolutional part.

B. Discussion

In this section, we provide the discussion of experimental results in line with the research question stated in Section I.

RQ1. Analysis of the alteration of recommendation lists in *TAaMR*. The first research question verifies whether the application of targeted adversarial attacks against images of low recommended products alters the behavior of a multimedia recommender. Table II shows the results of $CHR@N$ in source-target misclassification scenarios. We bold the best result for each attack scenario.

Starting with the analysis of the budget perturbation impact (ϵ), our results confirm that more powerful adversarial capabilities in perturbing input images have profound repercussions on the recommendation lists. For instance, $CHR@100$ of PGD attacks on *Sock→Running Shoes* is almost doubled between low budget ($\epsilon = 2$, $CHR@100 = 3.654$) and large budget ($\epsilon = 8, CHR@100 = 6.402$) perturbations. We may justify the mentioned results by looking at the targeted success rate shown in Table III. As expected, an increase in the budget is followed by clear improvements in the success of misclassifying the experimented CNN (i.e., ResNet50).

Distinguishing attack strategies, we recognize FGSM as lesser powerful than PGD. This effect depends on the lower success rate of the first attack strategy with respect to the second. For instance, when FGSM reaches the biggest misclassification rate (56.46%, FGSM attack on Maillot→Brassiere) the $CHR@100$ is one third of the corresponding PGD. (1.538 and 4.421, respectively).

Comparing the results between *TAaMR* scenarios with semantically similar/dissimilar source-target categories, we notice that the proposed approach is more efficient with the semantic closeness of source and target classes. Results in Table II confirm this trend in almost each experimental configuration. Moreover, we evidence that $CHR@100$ gets even worse when $\epsilon \leq 4$, motivated by the difficulties in performing the correct targeted misclassification (see Table III).

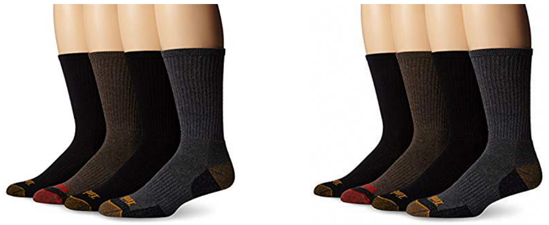
Finally, we compare *TAaMR* results regarding the recommender models (i.e., VBPR, and AMR). Table II clearly shows that the integration of the adversarial regularizer (see Section IV-A3) makes AMR less affected by the attacks compared to VBPR, but it is not completely safe. A plausible explanation is that *TAaMR* is still powerful because the rationale of AMR is to protect towards untargeted attacks instead of targeted ones. Finally, it is interesting to observe that the $CHR@100$ on AMR recommendation lists seems to reduce instead of increasing after the attack. This aspect opens further studies to interpret the consequence of the adversarial regularization on multimedia recommenders.

RQ2. Analysis of the visual appearance of perturbed products images.

TAaMR is valid as long as the attack does not produce *evident* artifacts on the item photos shown to the users. We measured the distortion between *original* and *perturbed* images objectively: (i) with traditional quality metrics (i.e., PSNR), (ii) emulating the way the user perceives the scene (SSIM) and (iii) understanding the semantic content (PMS) (see Section IV-A4). Table IV shows metric results calculated on attacked images.

TABLE IV: Average visual-quality metrics.

Attack	Amazon Men				Amazon Women			
	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Peak Signal-To-Noise Ratio (PSNR)								
FSGM	41.417	40.915	39.916	37.075	40.343	39.969	39.201	**36.770**
PGD	41.417	41.259	40.891	40.034	40.343	40.223	39.930	39.228
Structural Similarity Index (SSIM)								
FSGM	0.9926	0.9921	**0.9902**	0.9802	0.9935	0.9929	0.9912	0.9818
PGD	0.9926	0.9924	0.9920	0.9908	0.9935	0.9933	0.9929	0.9918
Perceptual Similarity Metric (PSM)								
FSGM	0.0132	0.0248	0.0397	0.0502	0.0066	0.0136	0.0235	0.0300
PGD	0.0328	0.0903	0.1877	**0.2368**	0.0175	0.0513	0.1246	0.2341



(a) original (sock)
probability: 60%
rec. position: 180th

(b) attacked (running shoe)
probability: 100%
rec. position: 14th

Fig. 2: Example of a product image before (a) and after (b) a PGD attack ($\epsilon = 8$) against VBPR on Amazon Men.

Examining the budget perturbation ϵ, we notice that the image distortion increases when the attack becomes stronger (i.e., the value of ϵ increases). However, even the worse metric values stay within their respective best ranges (e.g., the worst PSNR is 36.770 on Amazon Men dataset, with $\epsilon = 16$, the worst SSIM is 0.9902 on Amazon Men, with $\epsilon = 8$, and the worst PSM is 0.2368 on Amazon Men, with $\epsilon = 16$).

Furthermore, we compare FGSM and PGD on the two datasets. According to the PSNR and the SSIM, the two attacks perform almost equally from a visual point of view, with slight improvements shown by PGD (on average, 1.025 times better on the PSNR and 1.003 times better on the SSIM). Not surprisingly, this trend is inverted on PSM, where FGSM outperforms on average 2.059 times the iterative attack (i.e., PGD). We justify the previous analysis by the fact that PGD is way better than FGSM at generating attacked images able to fool the CNN (see Table III).

In conclusion, the visual analysis completes the *TAaMR* evaluation, by confirming that not only the proposed approach fooled the MR (and CNN), but the human customer (the main actor in any product/service provider) is also not aware that received recommendations, and item photos, have been maliciously tampered by an adversary. Fig. 2 is a real example generated during the experimented attack.

V. Related Work

A. Multimedia Recommender Systems

RS are categorized [39] as content-based filtering, collaborative filtering, and hybrid. The first class recommends items based on their content features [40]. The second category tracks users' interactions (e.g., ratings, reviews) to generate recommendations based on the assumptions that a user might like products already interacted by users with similar preferences [41]. The last category accomplishes the recommendation task integrating content data (e.g., knowledge graphs [42], users' reviews [43], images [11]) together with users' collaborative information [44]. Multimedia recommender systems fall inside the aforementioned third category. The basic intuition is to use multimedia features to improve recommendation performance. For instance, VBPR [12], [45] integrates deep visual features of product images. While, the authors of [13], [46] employ music and video features, respectively, in MR.

B. Security of Recommender Systems

Recommender models could be the victim of malicious users. The exploration of the security of RS started in the early 2000s with a deep interest in studying the injection of meticulously hand-engineered [47], [48], or machine-learning optimized [49], [50], poisoning profiles from multiple perspectives: altering the performance of collaborative RS for *pushing*, or *nuking* a specific item or set of items [15], detecting shilling profiles [51] and strengthening recommendation algorithms [52]. Recently, novel efforts have been dedicated to the study of the security of RS under the point of view of adversarial machine learning attacks (and defense) on recommender models [18]. The main research direction has been established in [19]. The authors propose an adversarial training strategy to limit the performance degradation caused by adversarial perturbation on users and item embeddings of a matrix factorization recommender models. Improvements in both robustness and recommendation make the proposed adversarial training studied, and implemented, in several models and domains (e.g., tensor-factorization [53], DNN-RS [54], [55], review-based recommendation [56]). Among these, Adversarial Multimedia Recommendation (AMR) [20] is the first work to investigate the worsening of recommendation performance of MR trained on untargeted adversarial examples.

VI. Conclusion and Future Work

In this work, we have explored the application of targeted adversarial attacks on input images for multimedia recommenders. The proposed approach, called *TAaMR* (evaluated on two real-world datasets) demonstrated the possibility of altering recommendations such that a low recommended product category could become three times more recommended by perturbing product images in a human-imperceptible way. Additionally, we have verified that the multimedia model using adversarial training slightly reduces the effectiveness of attacks. This open to novel solutions that we plan to explore. We aim to design a finer-grained visual attack to address a single item even within the same category (e.g., one kind of sock against another one). Furthermore, we intend to extend *TAaMR* integrating novel adversarial attacks (e.g., optimized to misuse the recommender) and evaluating the impact of defense strategies (e.g., adversarial training and defensive distillation) –to make the feature extraction more robust– as well as user-driven online evaluation with subjective visual metrics.

Acknowledgment

The authors acknowledge partial support of the following projects: Innonetwork CONTACT, Innonetwork APOLLON, ARS01_00821 FLET4.0, Fincons Smart Digital Solutions for the Creative Industry, PON OK-INSAID.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *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*, pp. 1106–1114, 2012.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 91–99, 2015.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Z. Yuan, Y. Lu, Z. Wang, and Y. Xue, "Droid-sec: deep learning in android malware detection," in *ACM SIGCOMM 2014 Conference, SIGCOMM'14, Chicago, IL, USA, August 17-22, 2014*, pp. 371–372, 2014.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[6] I. J. Goodfellow, P. D. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Commun. ACM*, vol. 61, no. 7, pp. 56–66, 2018.

[7] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pp. 372–387, 2016.

[8] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57, 2017.

[9] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, 2003.

[10] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Trans. Management Inf. Syst.*, vol. 6, no. 4, pp. 13:1–13:19, 2016.

[11] R. He and J. J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pp. 507–517, 2016.

[12] R. He and J. J. McAuley, "VBPR: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 144–150, 2016.

[13] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," *J. Data Semantics*, vol. 5, no. 2, pp. 99–113, 2016.

[14] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2643–2651, 2013.

[15] S. K. Lam and J. Riedl, "Shilling recommender systems for fun and profit," in *WWW*, pp. 393–402, ACM, 2004.

[16] I. Gunes, C. Kaleli, A. Bilge, and H. Polat, "Shilling attacks against recommender systems: a comprehensive survey," *Artif. Intell. Rev.*, vol. 42, no. 4, pp. 767–799, 2014.

[17] G. Yang, N. Z. Gong, and Y. Cai, "Fake co-visitation injection attacks to recommender systems," in *24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, February 26 - March 1, 2017*, 2017.

[18] Y. Deldjoo, T. D. Noia, and F. A. Merra, "Adversarial machine learning in recommender systems (aml-recsys)," in *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pp. 869–872, 2020.

[19] X. He, Z. He, X. Du, and T. Chua, "Adversarial personalized ranking for recommendation," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 355–364, 2018.

[20] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, and T. Chua, "Adversarial training towards robust multimedia recommender system," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

[21] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *CoRR*, vol. abs/1902.06705, 2019.

[22] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019.

[23] M. Deshpande and G. Karypis, "Item-based top-*N* recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, 2004.

[24] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[25] J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pp. 43–52, 2015.

[26] K. Grauman, "Computer vision for fashion: From individual recommendations to world-wide trends," in *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, p. 3, 2020.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[28] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[29] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 274–283, 2018.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[31] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

[32] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pp. 452–461, 2009.

[33] Y. Koren and R. M. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, pp. 77–118, Springer, 2015.

[34] K. Thung and P. Raveendran, "A survey of image quality measures," in *2009 International Conference for Technical Postgraduates (TECHPOS)*, pp. 1–4, Dec 2009.

[35] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, pp. 660–668, Sep. 2008.

[36] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[37] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pp. 694–711, 2016.

[38] K. He, X. Zhang, S. Ren, and J. 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*, pp. 770–778, 2016.

[39] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., *Recommender Systems Handbook*. Springer, 2011.

[40] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pp. 263–272, 2008.

[41] D. Goldberg, D. A. Nichols, B. M. Oki, and D. B. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[42] V. W. Anelli, T. Di Noia, E. Di Sciascio, A. Ragone, and J. Trotta, "How to make latent factors interpretable by feeding factorization machines with knowledge graphs," in *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, pp. 38–56, 2019.

[43] L. Chen, G. Chen, and F. Wang, "Recommender systems based on user reviews: the state of the art," *User Model. User-Adapt. Interact.*, vol. 25, no. 2, pp. 99–154, 2015.

[44] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, pp. 73–105, Springer, 2011.

[45] X. Geng, H. Zhang, J. Bian, and T. Chua, "Learning image and user features for recommendation in social networks," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 4274–4282, 2015.

[46] J. Donaldson, "A hybrid social-acoustic recommendation system for popular music," in *Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys 2007, Minneapolis, MN, USA, October 19-20, 2007*, pp. 187–190, 2007.

[47] Y. Deldjoo, T. D. Noia, and F. A. Merra, "Assessing the impact of a user-item collaborative attack on class of users," in *Proceedings of the 1st Workshop on the Impact of Recommender Systems co-located with 13th ACM Conference on Recommender Systems, ImpactRS@RecSys 2019), Copenhagen, Denmark, September 19, 2019*, 2019.

[48] V. W. Anelli, Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. A. Merra, "Sasha: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs," in *Proceedings of the 17th Conference on Extended Semantic Web Conference*, Springer, 2020.

[49] M. Fang, N. Z. Gong, and J. Liu, "Influence function based data poisoning attacks to top-n recommender systems," *CoRR*, vol. abs/2002.08025, 2020.

[50] M. Fang, G. Yang, N. Z. Gong, and J. Liu, "Poisoning attacks to graph-based recommender systems," in *Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC 2018, San Juan, PR, USA, December 03-07, 2018*, pp. 381–392, 2018.

[51] R. Bhaumik, C. Williams, B. Mobasher, and R. Burke, "Securing collaborative filtering against malicious attacks through anomaly detection," in *Proceedings of the 4th Workshop on Intelligent Techniques for Web Personalization (ITWP'06), Boston*, vol. 6, p. 10, 2006.

[52] M. P. O'Mahony, N. J. Hurley, N. Kushmerick, and G. C. M. Silvestre, "Collaborative recommendation: A robustness analysis," *ACM Trans. Internet Techn.*, vol. 4, no. 4, pp. 344–377, 2004.

[53] H. Chen and J. Li, "Adversarial tensor factorization for context-aware recommendation," in *RecSys*, pp. 363–367, ACM, 2019.

[54] F. Yuan, L. Yao, and B. Benatallah, "Adversarial collaborative neural network for robust recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pp. 1065–1068, 2019.

[55] Y. Du, M. Fang, J. Yi, C. Xu, J. Cheng, and D. Tao, "Enhancing the robustness of neural collaborative filtering systems under malicious attacks," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 555–565, 2019.

[56] D. Rafailidis and F. Crestani, "Adversarial training for review-based recommendations," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.*, pp. 1057–1060, 2019.