Assessing Perceptual and Recommendation Mutation of Adversarially-Poisoned Visual Recommenders

Vito Walter Anelli[a], Tommaso Di Noia[a], Daniele Malitesta[a] and Felice Antonio Merra[a]

[a]Politecnico di Bari, Via Edoardo Orabona, 4, 70126 Bari BA, Italy

Abstract

Visually-aware recommendation leverages visual signals of product images extracted through Deep Neural Networks to improve the recommendation performance. However, human-imperceptible adversarial noise can alter recommendation outcomes, e.g., pushing/nuking specific product categories. In this work, we provide 24 combinations of attack/defense strategies, and visual-based recommenders to 1) access performance alteration on recommendation and 2) empirically verify the effect on final users through offline-visual metrics. The results suggest defense is not protecting recommender models as expected, and shed light on the importance of human evaluation to identify visual attacks on recommendations. Source code, data, and experimental parameters are available at https://github.com/sisinflab/Perceptual-Rec-Mutation-of-Adv-VRs.

Keywords

Adversarial Machine Learning, Recommender Systems, Data Poisoning

1. Introduction

Recommender Systems (RSs) provide the set of the most relevant products to the customers of online sellers. In domains such as fashion and food, visual signals associated with pictures influence users' decisions. Benefiting from the power of Deep Neural Networks (DNNs) in extracting high-level visual aspects from images, the class of Visual Recommender Systems (VRSs) achieved significant success in learning high-quality recommendations. He and McAuley [1, 2] proposed Visual Bayesian Personalized Ranking (VBPR) demonstrating terrific performance improvement compared to BPR-MF by Rendle et al. [3] with the simple integration of image features extracted from AlexNet [4].

Unfortunately, DNNs are vulnerable to adversarial examples [5, 6] minimal-corrupted images crafted to fool the network. Szegedy et al. [5] formalized the adversarial generation problem by solving a box-constrained L-BFGS. Goodfellow et al. [7] used the sign of the gradient of the loss function to perturb the images in the Fast Gradient Sign Method (FGSM). Madry et al. [8] adapted FGSM and Basic Iterative Method [9] to *iteratively* update the perturbation and get stronger adversarial samples. Carlini and Wagner [10] (C & W) boosted the Szegedy et al. [5] strategy to craft powerful samples able to deceiving state-of-the-art adversarial detector [11]. However, the

AIxIA 2020 Discussion Papers Workshop
vitowalter.anelli@poliba.it (V. W. Anelli); tommaso.dinoia@poliba.it (T. Di Noia); daniele.malitesta@poliba.it (D. Malitesta); felice.merra@poliba.it (F. A. Merra)

a. Clean
Rec. Position: 68th

b. Attack + **T**
Rec. Position: 10th
LPIPS: 0.5484

c. Attack + **AT**
Rec. Position: 27th
LPIPS: 0.5347

d. Attack + **FAT**
Rec. Position: 40th
LPIPS: 0.3447

Figure 1: (a) is the image of a *low-recommended* product. (b, c, d) are the perturbed versions with PGD ($\epsilon = 8$) applied against DNNs without defense (T), or with the Adversarial Training (AT) and Free AT (FAT). The attacks have pushed the product towards *higher* ranking positions without *visually-perceptible* artifacts.

Adversarial Training, proposed by Goodfellow et al. [7], has demonstrated substantial DNN's protection when adversarial samples are injected into the training data at a long-time training cost. This issue has been recently addressed by Shafahi et al. [12] with the proposal of the $3 - 30$ times faster Free Adversarial Training.

Consequently, adversarially-perturbed product images have been also shown to fool the DNNs used in VRS to extract the visual features [13]. Tang et al. [14] tested the accuracy degradation when VBPR is trained on noisy images (integrity attack), while Di Noia et al. [15] demonstrated the adversary's capability to increase (or decrease) the recommendability of a category of products (integrity attack) even on the *adversarial regularized* [16] version of VBPR, namely AMR [14].

In this paper, we investigate the efficacy of defensive mechanisms [7, 12] against powerful attacks [7, 8, 17] when the adversary wants to alter the recommendation lists of a VRS by *poisoning* the training data by inserting adversarial product images, e.g., one perturbs images of low popular products so that they are misclassified as popular ones. Furthermore, we provide a visual-oriented evaluation of adversarial images through offline *visual metrics trying to mimicking human evaluation* to verify to what extent users might become aware of such subtle data poisoning in the received recommendations (Figure 1).

The main contributions of this paper are twofold: (1) we verify the inefficacy of state-of-the-art adversarial training procedure in defending the DNNs used in VRS from *adversarially-poisoned training product images*; (2) we evaluate the human-perceptibility with offline measures.

2. The Threat Model

Given the set of users $\mathcal{U}$, items $\mathcal{I}$, the matrix of historical interactions $\mathcal{S}$, the recommendation problem is defined as the task to suggest products by maximizing the user's gain $g(u)$. The state-of-the-art RS, BPR-MF [3, 18], solves $g(u)$ by maximizing a loss function over a set of triplets $\mathcal{T}$ defined as:

$$\mathcal{L}_{BPR} = \sum_{(u,i,j)\in\mathcal{T}} -\ln\sigma(\hat{s}_{ui} - \hat{s}_{uj}) + \lambda\|\theta\|_2^2 \tag{1}$$

where λ is the regularization coefficient, $\sigma(\cdot)$ is a sigmoidal function, and $\hat{s}_{ui}$, the predicted preference score of the user u on the item i measured as $\hat{s}_{ui} = p_u^T q_i$. Here, p_u and q_i are the user-specific and item-specific latent features, respectively. Then, for each item i, x_i is the associated product image. Let f_i the visual signal extracted from a DNN whose function model is F, i.e., f_i is the output of the first fully-connected layer placed immediately after the convolutional part. Then, He and McAuley [1] extended BPR-MF by integrating the visual signal while measuring $\hat{s}_{ui}$. The new formulation is:

$$\hat{s}_{ui} = p_u^T q_i + \underbrace{\rho_u^T(\mathbf{E}f_i)}_{\textit{visual signal}} \tag{2}$$

where ρ_u is the user's visual factor, and $\mathbf{E}$ is an embedding matrix to project f_i into the same dimensional space as for ρ_u.

The dependence of a VRS from the visual signal in Equation 2 has been exploited by adversaries to poison the training data with the insertion of adversarial samples [14, 15, 19]. To generate the **targeted** adversarial attack the optimization problem formulation is:

$$\max_{\delta_i:\|\delta_i\|_p\leq\epsilon} \mathcal{L}_F(x_i + \delta_i, y_i) \text{ s.t. } y_i = m \tag{3}$$

where $\mathcal{L}_F$ is the cost function of F, δ_i is the ϵ-bounded perturbation of x_i that will make the product image be misclassified by F as the (more popular) product category m, and $\|\cdot\|_p$ is the L_p norm. For instance, the adversary can poison the data adding a perturbed image of "*Jersey, T-shirt*" misclassified as "*Brassiere*" (Fig. 1) causing a variation in the VRS since f_i will be extracted from $x_i^{adv} = x_i + \delta_i$.

Recently, studies on the robustification of DNNs have shown the adversarial training by Goodfellow et al. [7] is one of the most prominent defense technique. After the definition of the adversary threat model (i.e., the attack strategy), the adversarial minimax formulation is:

$$\min_{\widetilde{\theta}} \sum_{(x_i,y_i)\in\mathcal{I}} \max_{\delta_i:\|\delta_i\|_p\leq\epsilon} \mathcal{L}_F(x_i + \delta_i,\ y_i) \tag{4}$$

where $\widetilde{\theta}$ represents the model parameters of the robustified network ($\widetilde{F}$).

Let $\widetilde{f}_i$ the visual features of the image x_i associated to a product image extracted from $\widetilde{F}$. In this work, we want to verify if the application of adversarial training methods can limit poisoning attacks against VRSs [14, 15] since each user-item score prediction $\hat{s}_{ui}$ depends on $\widetilde{f}_i$. Furthermore, we want to investigate whether the usage of adversarial trained DNNs will make the adversarial perturbation evident to such an extent that it makes the perturbed samples identifiable via a human evaluation.

Table 1
$CHR@20$ results on `Amazon Women` and `Amazon Men`. We mark in **bold** the most effective attacks.

Model	Attack	Amazon Women			Amazon Men		
		T	AT	FAT	T	AT	FAT
VBPR	No-Attack	*0.4377*	*0.5108*	*0.3417*	*0.6352*	*0.3028*	*0.3702*
	FGSM ($\epsilon = 4$)	0.3860	0.6032	0.6088	0.5665	0.6029	0.5688
	FGSM ($\epsilon = 8$)	0.4057	0.6186	**0.6313**	0.6052	0.5879	0.5596
	PGD ($\epsilon = 4$)	0.4377	0.6309	0.6263	1.0936	0.6211	0.5778
	PGD ($\epsilon = 8$)	**1.4462**	**0.6413**	0.6139	**1.5736**	0.6247	0.6141
	C&W	0.4147	0.6280	0.5729	0.5972	**0.6652**	**0.6444**
AMR	No-Attack	*0.9449*	*0.8342*	*0.5063*	*0.3876*	*0.4924*	*0.1070*
	FGSM ($\epsilon = 4$)	**1.3173**	0.7135	0.4565	0.3295	0.4332	0.4103
	FGSM ($\epsilon = 8$)	1.2814	0.7137	0.4429	0.3053	0.4318	0.4007
	PGD ($\epsilon = 4$)	1.1958	0.6473	0.4900	0.8064	0.4435	0.4173
	PGD ($\epsilon = 8$)	1.2377	0.6770	0.4445	**2.1264**	0.4323	0.3942
	C&W	1.3012	0.7159	0.4977	0.3610	0.4293	**0.4378**

3. Experiments

Setup. The experiments are conducted on two fashion datasets, i.e., `Amazon Women` and `Amazon Men` made publicly available by He and McAuley [2]. They come with both users' ratings and product pictures uploaded by the platform owner and third-party sellers (say, the possible adversaries). `Amazon Women` counts 16668 users, 2981 items, and 54473 ratings, while `Amazon Men` counts 24379, 7371, and 89020. We split the data following the time-aware leave-one-out protocol [16].

To empirically study the efficacy of defenses and evaluate the visual appearance of adversarial samples, we tested two VRS: VBPR by He and McAuley [1], and AMR by Tang et al. [14], a VBPR extension that includes the adversarial regularizer of visual features proposed by He et al. [16]. The complete set of experimental parameters is reported in the GitHub repository.

Evaluation of Recommendation Performance. Table 1 shows the recommendation variation before and after the attacks. We evaluate the variation of recommendation with the $CHR@K$ [15], that measures the average number of a (pushed) category of items in the top-K recommendation lists. In particular, results in Table 1 are measured on the following source-target combinations: "*Sandal*"-"*Running Shoe*" for `Amazon Men`, while "*Jersey, T-shirt*"-"*Brassiere*" for `Amazon Women`, where the adversary tries to push the recommendability of a source category by perturbing the product picture to be classified as a target class, e.g., the class of a very popular category.

Analyzing VBPR outcomes, PGD attack shows the highest variation of $CHR@20$ in the defense-free experiments. For instance, PGD ($\epsilon = 8$) increases by more than 2.3 times the $CHR@20$ of the source category in the <`Amazon Women`, VBPR, Traditional> setting. The same trend is not true for the defense contexts. C&W attacks have increased the $CHR@20$ by 71.09%, while PGD ($\epsilon = 8$) by 69.35%. Furthermore, Table 1 confirms that the adversarial training strategies have failed in protecting VBPR since the data poisoning is always effective in any defended settings.

Table 2
Average values of Success Rate (SR), Feature Loss (FL) and Learned Perceptual Image Patch Similarity ($LPIPS$) for each <dataset, attack, defense> combination. $LPIPS$ is multiplied by 100. We mark in **bold** the best results for each considered metric.

Dataset	Attack	Image Feature Extractor								
		Traditional			Adversarial Training			Free Adversarial Training		
		SR	FL	$LPIPS$	SR	FL	$LPIPS$	SR	FL	$LPIPS$
Amazon Women	FGSM ($\epsilon = 4$)	17.70%	0.0096677	0.2388	0.00%	0.0000113	0.1353	0.00%	0.0000094	0.1041
	FGSM ($\epsilon = 8$)	28.32%	0.0220499	2.8505	2.65%	0.0000851	1.8298	0.00%	0.0000671	1.2119
	PGD ($\epsilon = 4$)	84.96%	0.0276645	**0.1860**	0.00%	0.0000119	0.1093	0.00%	0.0000102	0.0860
	PGD ($\epsilon = 8$)	**100.00%**	**0.1303309**	1.1136	3.54%	0.0000974	0.7683	0.00%	0.0000735	0.6369
	C & W	89.38%	0.0212380	0.2678	**6.19%**	**0.0001770**	**0.0731**	**6.19%**	**0.0003376**	**0.0816**
Amazon Men	FGSM ($\epsilon = 4$)	65.45%	0.0140948	0.1861	18.32%	0.0000330	0.1407	15.18%	0.0000278	0.1074
	FGSM ($\epsilon = 8$)	86.91%	0.0363190	1.7124	23.56%	0.0002658	2.2903	20.42%	0.0002320	1.2293
	PGD ($\epsilon = 4$)	96.86%	0.0368843	**0.1669**	18.32%	0.0000334	**0.1257**	15.18%	0.0000283	**0.0892**
	PGD ($\epsilon = 8$)	**100.00%**	**0.1349854**	0.6916	24.08%	0.0002801	0.7997	20.94%	0.0002371	0.6468
	C & W	89.01%	0.0205172	0.2279	**48.17%**	**0.0028022**	0.2688	**42.41%**	**0.0019080**	0.1490

Investigating AMR results, the attacks are quite effective in the defense-free settings as much as in VBPR, and confirm PGD ($\epsilon = 8$) as the most powerful method. Interestingly, the joint usage of (1) adversarial training procedures on the DNN and (2) the adversarial regularization on the recommender embeddings (APR) significantly reduced the effectiveness of the dataset poisoning. Indeed, 75% of attacks have not increased the $CHR@20$ of the low popular category of products.

Visual Evaluation. To investigate the efficacy of attacks in poisoning the VRS, we studied the attack Success Rate (SR), the Feature Loss (FL), and the Learned Perceptual Image Patch Similarity ($LPIPS$) [20]. Given the importance that visual features hold in VRSs, FL calculates the MSE between extracted features before and after the attack. That is, it provides a measure of visual features' shifting in the latent space, and how this has affected recommendation. The idea behind $LPIPS$ is to produce a perceptual distance value between two similar images by leveraging (1) knowledge extracted from convolutional layers inside state-of-the-art CNNs and (2) collected human visual judgments about those pairs of similar images. We computed this metric fine-tuning a VGG [21] network since Zhang et al. [20] proposed this configuration as the best one at imitating a real human-evaluation in circumstances comparable to visual attacks.

Table 2 reports the $LPIPS$ results, along with SR and FL values. It is worth recalling that a large (small) FL value stands for *semantically* different (similar) images from DNN's point of view. Similarly, a large (small) $LPIPS$ value means the two compared images would likely be considered as *visually* different (similar) by humans.

Two general observations arise here. First, the FL is strictly correlated to the SR, i.e., *an attack is successful when the extracted features are noticeably shifted in the latent space.* Second, all attack combinations are able to keep $LPIPS$ values within low ranges, in accordance with the *imperceptible* nature of adversarial perturbations on images [5]. Thus, we connect this obtained measure with the attack efficacy in both failing the classifier (i.e., the DNN) and the VRS. What follows is a detailed evaluation of scenarios involving —or not— defensive techniques for the DNN.

Defense-free Setting. In the defense-free scenario, PGD ($\epsilon = 4$) is the least perceptible attack

—with the lowest $LPIPS$ values— even considering a near-100% SR and a successful pushing of attacked products. On the other hand, FGSM ($\epsilon = 8$) fails to hide the produced perturbations, reaching the highest perceptible visual difference on `Amazon Women` (2.8505). Coherently, this setting also shows a low SR and a weak alteration of visual recommendations (see Table 1).

Defense Setting. Let us focus on the two defenses. Here, it becomes fundamental to consider the $LPIPS$ value along with its corresponding SR and recommendation variations. As a matter of fact, in a defense context, where all attacks averagely tend to perform worse at failing the DNN classifier, a measured low average $LPIPS$ value might trivially mean *very few* images were *successfully* attacked. For instance, the described situation occurs in the combination <`Amazon Men`, PGD ($\epsilon = 4$), Adversarial Training>. However, since these attacks have still been effective in *pushing* low ranked category products (as evident in Table 1), then adversaries could exploit their hardly-human perceptibility to craft even stronger perturbations (e.g., increasing ϵ). An intriguing situation is when $LPIPS$ on the defended DNN is higher than the non-defended one. The worst case is <`Amazon Men`, FGSM ($\epsilon = 8$), Adversarial Training>, which shows a 34% increase of $LPIPS$ compared to the Traditional training. We explain this result considering that and attack might need to produce larger perturbations to move the category of the few correctly attacked images (about 24% in the cited example) towards the targeted one. Not only is the attack inefficient, but it risks human identification.

4. Conclusion

We have presented an empirical study to evaluate the efficacy of defenses (i.e., Adversarial Training and Free Adversarial Training) to protect DNNs on top of visually-aware recommender systems when poisoning product image datasets with adversarial attacks. Experiments on state-of-the-art visual recommenders VBPR and AMR trained on two datasets (i.e., `Amazon Women` and `Amazon Men`) demonstrated the alarming weakness of adversarial training in protecting the recommendation performance. Furthermore, the visual evaluation suggested defense scenarios with few successfully attacked images and barely perceptible visual artifacts that still keep breaking recommendation performance are blind spots that adversaries could explore deeper for their malicious purposes. Conclusively, we plan to study attack efficacy on overall recommendation performance (accuracy and beyond-accuracy), propose novel end-to-end defenses, provide a parallel in-depth study on the impact of perturbed images for humans, the users of the platforms.

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] R. He, J. J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback, in: D. Schuurmans, M. P. Wellman (Eds.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, AAAI Press, 2016, pp. 144–150. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11914.

[2] R. He, J. J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 507–517. URL: https://doi.org/10.1145/2872427.2883037. doi:10.1145/2872427.2883037.

[3] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: bayesian personalized ranking from implicit feedback, in: J. A. Bilmes, A. Y. Ng (Eds.), UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009, AUAI Press, 2009, pp. 452–461. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1630&proceeding_id=25.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), 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, 2012, pp. 1106–1114. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL: http://arxiv.org/abs/1312.6199.

[6] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: H. Blockeel, K. Kersting, S. Nijssen, F. Zelezný (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III, volume 8190 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 387–402. URL: https://doi.org/10.1007/978-3-642-40994-3_25. doi:10.1007/978-3-642-40994-3_25.

[7] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6572.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, 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, OpenReview.net, 2018. URL: https://openreview.net/forum?id=rJzIBfZAb.

[9] A. Kurakin, I. J. Goodfellow, 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, OpenReview.net, 2017. URL: https://openreview.net/forum?id=HJGU3Rodl.

[10] N. Carlini, 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, IEEE Computer Society, 2017, pp. 39–57. URL: https://doi.org/10.1109/SP.2017.49. doi:10.1109/SP.2017.49.

[11] N. Carlini, D. A. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, A. Sinha (Eds.), Proceed-

ings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017, ACM, 2017, pp. 3–14. URL: https://doi.org/10.1145/3128572.3140444. doi:10.1145/3128572.3140444.

[12] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. P. Dickerson, C. Studer, L. S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, 2019, pp. 3353–3364. URL: http://papers.nips.cc/paper/8597-adversarial-training-for-free.

[13] Y. Deldjoo, T. D. Noia, F. A. Merra, A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks, CoRR abs/2005.10322 (2020). URL: https://arxiv.org/abs/2005.10322. arXiv:2005.10322.

[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. URL: https://doi.org/10.1109/TKDE.2019.2893638. doi:10.1109/TKDE.2019.2893638.

[15] T. Di Noia, D. Malitesta, F. A. Merra, Taamr: Targeted adversarial attack against multimedia recommender systems, in: 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN Workshops 2020, Valencia, Spain, June 29 - July 2, 2020, IEEE, 2020, pp. 1–8. URL: https://doi.org/10.1109/DSN-W50199.2020.00011. doi:10.1109/DSN-W50199.2020.00011.

[16] X. He, Z. He, X. Du, T. Chua, Adversarial personalized ranking for recommendation, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 355–364. URL: https://doi.org/10.1145/3209978.3209981. doi:10.1145/3209978.3209981.

[17] N. Carlini, D. A. Wagner, Defensive distillation is not robust to adversarial examples, CoRR abs/1607.04311 (2016). URL: http://arxiv.org/abs/1607.04311. arXiv:1607.04311.

[18] S. Rendle, W. Krichene, L. Zhang, J. R. Anderson, Neural collaborative filtering vs. matrix factorization revisited, in: R. L. T. Santos, L. B. Marinho, E. M. Daly, L. Chen, K. Falk, N. Koenigstein, E. S. de Moura (Eds.), RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020, ACM, 2020, pp. 240–248. URL: https://doi.org/10.1145/3383313.3412488. doi:10.1145/3383313.3412488.

[19] Z. Liu, M. A. Larson, Adversarial item promotion: Vulnerabilities at the core of top-n recommenders that use images to address cold start, CoRR abs/2006.01888 (2020). URL: https://arxiv.org/abs/2006.01888. arXiv:2006.01888.

[20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 586–595. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00068.

[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1409.1556.