V-Elliot: Design, Evaluate and Tune Visual Recommender Systems

VITO WALTER ANELLI, Politecnico di Bari, Italy
ALEJANDRO BELLOGÍN, Autónoma Madrid, Spain
ANTONIO FERRARA, Politecnico di Bari, Italy
DANIELE MALITESTA, Politecnico di Bari, Italy
FELICE ANTONIO MERRA, Politecnico di Bari, Italy
CLAUDIO POMO, Politecnico di Bari, Italy
FRANCESCO MARIA DONINI, Università della Tuscia, Italy
TOMMASO DI NOIA, Politecnico di Bari, Italy

The paper introduces Visual-Elliot (V-Elliot), a reproducibility framework for Visual Recommendation systems (VRSs) based on Elliot. framework provides the widest set of VRSs compared to other recommendation frameworks in the literature (i.e., 6 state-of-the-art models which have been commonly employed as baselines in recent works). The framework pipeline spans from the dataset preprocessing and item visual features loading to easily train and test complex combinations of visual models and evaluation settings. V-Elliot provides an extended set of features to ease the design, testing, and integration of novel VRSs into V-Elliot. The framework exploits of dataset filtering/splitting functions, 40 evaluation metrics, five hyper-parameter optimization methods, more than 50 recommendation algorithms, and two statistical hypothesis tests. The files of this demonstration are available at: github.com/sisinflab/elliot.

CCS Concepts: • **Information systems** → **Recommender systems**; *Collaborative filtering*.

Additional Key Words and Phrases: Visual recommendation, Recommender Systems, Reproducibility

ACM Reference Format:
Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. V-Elliot: Design, Evaluate and Tune Visual Recommender Systems. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460231.3478881

1 INTRODUCTION

Recommender systems (RSs) power most of today's online platforms by providing users with a tool to explore vast and heterogeneous catalogs of products through personalized lists of items [16]. In some domains, such as fashion [13], food [8], or tourism [24], the visual appearance of a product image (e.g., piece of clothing or dish) is crucially important since it may affect user's final decision [11, 12]. Visual Recommender Systems (VRSs) integrate visual features of product images extracted through an image feature extractor (referred to as IFE, usually a CNN) into the recommendation pipeline to learn more tailored user profiles, overcoming issues such as data sparsity and cold-start [12]. The business of several online platforms is based on user-generated products and images (e.g., Pinterest, Amazon, Zalando, and

Manuscript submitted to ACM

Table 1. Most popular Visual Recommender Systems from the literature. For each work, we report its reference, publication year, adopted side information (i.e., either the image or the extracted visual feature of the item), the image feature extractor (with the chosen extraction layer and the training strategy), and link to the official code (if any). FC: fully-connected, FM: feature maps.

VRS	Year	Side Info		Image Feature Extractor				Code
				Extraction Layer		Training		
		Image	*Feature*	*FC*	*FM*	*Pretrained*	*End-to-End*	
VBPR [12]	2016		✓	✓		✓		×
DeepStyle [17]	2017		✓	✓		✓		×
DVBPR [15]	2017	✓		✓			✓	**[link]**
ACF [4]	2017		✓		✓	✓		**[link]**
VNPR [19]	2018		✓	✓		✓		×
AMR [26]	2020		✓	✓		✓		**[link]**

Instagram). Consequently, Academia and Industry have channelled a considerable effort in designing novel approaches for visual recommendation [5, 14, 20]. Table 1 provides an outline of the most popular VRSs adopted as baselines in the recent literature, with some technical information on the input data type, the extraction layer and the training methodology for the IFE, and the official code link (if available).

Despite their adoption as baselines in several recent works (e.g., [2, 9, 27–30]), to date nobody provided a unique framework implementing all these VRSs. Moreover, oftentimes, reproducibility is not even a feasible option since an official code is not always released (see "Code" in Table 1). Parra et al. [21] have recently proposed a tutorial on visual recommendation, presenting some (but not all) the above cited VRSs. Nevertheless, their work was not devoted to integrate the visual models into a complete framework for recommendation. Additionally, the copiousness of novel recommendation algorithms has generated confusion about choosing the correct baselines, the hyperparameter optimization, and the experimental evaluation to follow [22, 23]. Unreproducible evaluation and unfair comparisons [25] have recently arisen as a critical issue in the recommender systems community [6]. To this end, Anelli et al. [1] proposed Elliot, a framework for rigorous and reproducible recommender systems. The project is publicly available on GitHub, and provides several strategies for dataset loading, prefiltering, and splitting, along with hyperparameter optimization, recommendation models, and statistical hypothesis tests to build a reproducible experimental benchmark.

This work aims to provide a comprehensive demonstration of how to use Elliot for visual-based recommendation. Elliot for Visual recommendation (V-Elliot) implements all 6 VRSs from Table 1, with the possibility of leveraging: *(i)* a wide range of visual side information as input (e.g., the product image or its high-level visual feature extracted through a CNN-based IFE), *(ii)* a specific data input pipeline (implemented in `TensorFlow`) to efficiently handle memory-intensive streams of multidimensional data and inject them seamlessly into the recommendation flow, and *(iii)* an easy-to-use tool to train and test complex configurations of heterogeneous state-of-the-art (and custom) recommender systems by combining V-Elliot with the Elliot environment[1].

2 V-ELLIOT: THE VISUAL RECOMMENDATION FRAMEWORK

Elliot for Visual recommendation (V-Elliot) executes complex and reproducible experimental flows. As pointed out in Anelli et al. [1], the flexibility of the framework allows the user to design and run multiple possible settings through a concise configuration file, while seven modules are transparently loaded, each playing a specific functional role in the experimental flow. In addition to the already-existing modules (Figure 1), V-Elliot introduces a component to handle the loading and injection of visual side information (e.g., images and visual features) into the recommendation model.

[1]The code, the data, and the configuration files of this demonstration are publicly available at: https://github.com/sisinflab/elliot.

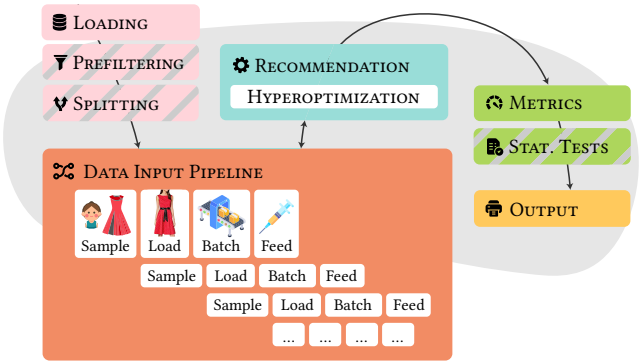

Fig. 1. Overview of V-Elliot. After the initial Loading (optionally complemented by Prefiltering and Splitting strategies), the Data Input Pipeline interacts with the Recommendation module to inject the visual data and train the model. The Metrics module evaluates the performance, whose values can be validated by statistical hypothesis tests. The Output module reports statistics and results.

Table 2. Measured accuracy and beyond-accuracy metrics for the tested Visual Recommender Systems and datasets on top-100 recommendation lists. Best values are reported in **bold**, while the second-best are underlined.

Model	Accuracy				Beyond-Accuracy				
	HR	*nDCG*	*Prec*	*MAP*	*EFD*	*EPC*	*Gini*	*SE*	*iCov*
`Amazon Baby` - [Elliot Configuration File: demo_amazon_baby.yml]									
VBPR	.0743	.0160	.0007	.0008	.0075	.0008	.5730	10.0093	1386
DVBPR	.0413	.0082	.0004	.0004	.0039	.0004	.1421	7.7011	370
ACF	**.1221**	**.0384**	**.0012**	**.0022**	**.0169**	**.0018**	.6711	10.1862	1392
DeepStyle	.0561	.0117	.0006	.0005	.0055	.0006	**.7490**	**10.2992**	**1393**
VNPR	.0479	.0112	.0005	.0006	.0052	.0005	.2058	8.5920	965
AMR	.0858	.0192	.0009	.0010	.0092	.0009	.5752	10.0083	1389
`Amazon Boys & Girls` - [Elliot Configuration File: demo_amazon_boys_girls.yml]									
VBPR	.0295	.0068	.0003	.0004	.0036	.0003	.4141	11.0699	3687
DVBPR	.0309	.0082	.0003	.0005	.0038	.0004	.4692	11.2376	3842
ACF	.0351	.0075	.0004	.0003	.0037	.0004	.0257	6.7975	120
DeepStyle	**.0653**	**.0210**	**.0007**	**.0013**	**.0099**	**.0010**	.2886	10.5499	3176
VNPR	.0260	.0053	.0003	.0003	.0029	.0003	**.6421**	**11.6361**	**3925**
AMR	.0365	.0094	.0004	.0005	.0047	.0004	.5349	11.4075	3902

The **Loading** module already supports various information sources (e.g., item features, semantic information [3], visual embeddings [12], and images [15]). As for the visual-based input data, the user can indicate the folder path where images (or features) are stored in separate files, which will be later injected *on-the-fly* into the framework when necessary. It is common knowledge that this strategy could alleviate the impact of memory-intensive experiments involving multidimensional visual data, which rarely can be pre-loaded into memory in advance. Users can also configure the settings for data pre-processing. In this respect, the **Prefiltering** module offers, among all, the possibility of applying the *filter-by-rating* and *k-core* strategies on the data, where the former removes user-item interactions whose preference score is smaller than a fixed (or data-based) threshold, and the latter filters out users, items, or both, with less than k interactions. Interestingly, the implementation of *k-core* algorithm also allows to retain cold users and items. Then, the **Splitting** module provides various temporal- and random-based splitting strategies, ranging from hold-out to cross-validation mechanisms. V-Elliot leverages a **Data Input Pipeline** to efficiently load visual-based input data and feed VRSs with it. The module is built upon the popular `TensorFlow` data input pipeline, which operates according to the producer/consumer paradigm, and consists of the following steps: *(i)* the next user-item interaction is sampled from the training set, *(ii)* visual data that has to be associated with the sample is loaded and (optionally) pre-processed,

e.g., undergoing a normalization phase, *(iii)* samples are (optionally) grouped into batches, and *(iv)* the batches feed the recommendation algorithm. The **Recommendation** module interacts with the Data Input Pipeline, and integrates with an ever-growing set of state-of-the-art recommendation models seamlessly. To the best of our knowledge, V-Elliot is the framework providing the highest number of VRSs from the literature integrated into a complete system for recommendation (see again Table 1). Moreover, the simplicity of extending the set of available recommender systems through custom and external models, and an exhaustive number of hyper-parameter tuning strategies considerably ease the prototyping phase. The training procedure is assisted by the **Metrics** module that evaluates the model performance (with metrics ranging from accuracy to beyond-accuracy ones) and drives the selection of the best hyper-parameters configuration. Furthermore, the V-Elliot memory-optimized version of the visual-based Data Input Pipeline is also exploited to speed up the evaluation process. The evaluation phase may be further refined by computing two statistical hypothesis tests, i.e., *Wilcoxon* and *Paired t-test*, using the **Statistical Tests** module. Finally, V-Elliot collects the results through the **Output** module, which stores detailed performance tables, whereas model weights and recommendation lists may be saved for the sake of reproducibility, further analysis, and future experiments.

3 EXECUTION OF AN EXPERIMENTAL FLOW

Setting. To encourage researchers to try V-Elliot, we show the experiments run on two fashion datasets (i.e., `Amazon Baby` and `Amazon Boys & Girls` [11, 18]) filtered through the 5-core technique as suggested in He and McAuley [11, 12]. The final statistics are: 606 users, 1761 items, and 3882 interactions for `Amazon Baby`, and 1425 users, 5019 items, and 9213 interactions for `Amazon Boys & Girls`. For each item image, we have extracted high-level visual features with a pre-trained ResNet50 [10], following the findings shown in [7]. We split the data adopting the temporal leave-one-out protocol. To tune the hyper-parameters on the validation set, we performed a grid search using *HR*@100 as the validation metric. Table 2 presents the accuracy and beyond-accuracy metric values measured on the top-100 recommendation lists for each best model. The Elliot configuration files are reported in Table 2.

Results. Table 2 shows that ACF is the most accurate model on `Amazon Baby`, providing also the most novel recommendation lists (i.e., *EFD* and *EPC*) and being the second-to-best regarding diversity and coverage (i.e., *Gini*, *SE*, and *iCov*). Interestingly, DeepStyle settles as one of the most accurate models on `Amazon Boys & Girls`. However, ACF still reaches remarkable accuracy results (it is the third-best recommender), confirming the performance observed on `Amazon Baby`. It is worth mentioning that the proposed analysis could be easily extended to wider search spaces, more metrics (e.g., bias measures), and additional (non-visual) recommender models (e.g., deep neural collaborative models), to eventually build an exhaustive evaluation workflow for recommendation.

4 CONCLUSION

Visual recommendation lacks comprehensive benchmarks, thus making unreproducible evaluation and unfair comparison a major concern. This demonstration presents a reproducible visual recommendation framework (V-Elliot) reviewing, implementing, and integrating six popular visual recommenders in the recent literature. V-Elliot takes advantage of all the features available in Elliot, such as data splitting and filtering, hyperparameters optimization, and evaluation strategies together with a GPU-optimized data input pipeline. The demonstration proceeds by showing the execution of a (reproducible) benchmark where a set of visual models is tested on two well-known datasets.

ACKNOWLEDGMENTS

The authors acknowledge partial support of PID2019-108965GB-I00, 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, PIA ERP4.0, and IPZS-PRJ4_IA_NORMATIVO.

REFERENCES

[1] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR*. ACM, 2405–2414.

[2] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images. In *SIGIR*. ACM, 1094–1103.

[3] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Azzurra Ragone, and Joseph Trotta. 2019. How to Make Latent Factors Interpretable by Feeding Factorization Machines with Knowledge Graphs. In *ISWC (1) (Lecture Notes in Computer Science, Vol. 11778)*. Springer, 38–56.

[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. ACM, 335–344.

[5] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *KDD*. ACM, 2662–2670.

[6] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *RecSys*. ACM, 101–109.

[7] Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A Study on the Relative Importance of Convolutional Neural Networks in Visually-Aware Recommender Systems. In *CVPR Workshops*. 3961–3967.

[8] David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting Food Choice Biases for Healthier Recipe Recommendation. In *SIGIR*. ACM, 575–584.

[9] Francesco Gelli, Tiberio Uricchio, Xiangnan He, Alberto Del Bimbo, and Tat-Seng Chua. 2020. Learning Visual Elements of Images for Discovery of Brand Posts. *ACM Trans. Multim. Comput. Commun. Appl.* 16, 2 (2020), 56:1–56:21.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society.

[11] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW*. ACM, 507–517.

[12] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. AAAI Press, 144–150.

[13] Yang Hu, Xi Yi, and Larry S. Davis. 2015. Collaborative Fashion Recommendation: A Functional Tensor Factorization Approach. In *ACM Multimedia*.

[14] Shatha Jaradat, Nima Dokoohaki, Humberto Jesús Corona Pampín, and Reza Shirvany. 2020. Second Workshop on Recommender Systems in Fashion - fashionXrecsys2020. In *RecSys*. ACM, 632–634.

[15] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *ICDM*. IEEE Computer Society, 207–216.

[16] Walid Krichene and Steffen Rendle. 2020. On Sampled Metrics for Item Recommendation. In *KDD*. ACM, 1748–1757.

[17] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *SIGIR*. ACM, 841–844.

[18] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR*. ACM, 43–52.

[19] Wei Niu, James Caverlee, and Haokai Lu. 2018. Neural Personalized Ranking for Image Recommendation. In *WSDM*. ACM, 423–431.

[20] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *KDD*. ACM, 2311–2320.

[21] Denis Parra, Antonio Ossa-Guerra, Manuel Cartagena, Patricio Cerda-Mardini, and Felipe del-Rio. 2021. VisRec: A Hands-on Tutorial on Deep Learning for Visual Recommender Systems. In *IUI Companion*. ACM, 5–6.

[22] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *RecSys*. ACM, 129–136.

[23] Alan Said and Alejandro Bellogín. 2014. Rival: a toolkit to foster reproducibility in recommender system evaluation. In *RecSys*. ACM, 371–372.

[24] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. 2020. PicTouRe - A Picture-Based Tourism Recommender. In *RecSys*. ACM, 597–599.

[25] Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. 2020. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In *RecSys*. ACM, 23–32.

[26] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2020. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Trans. Knowl. Data Eng.* 32, 5 (2020), 855–867. https://doi.org/10.1109/TKDE.2019.2893638

[27] Le Wu, Lei Chen, Richang Hong, Yanjie Fu, Xing Xie, and Meng Wang. 2020. A Hierarchical Attention Model for Social Contextual Image Recommendation. *IEEE Trans. Knowl. Data Eng.* 32, 10 (2020), 1854–1867.

[28] Ruiping Yin, Kan Li, Jie Lu, and Guangquan Zhang. 2019. Enhancing Fashion Recommendation with Visual Compatibility Relationship. In *WWW*. ACM, 3434–3440.

[29] Wenhui Yu, Xiangnan He, Jian Pei, Xu Chen, Li Xiong, Jinfei Liu, and Zheng Qin. 2019. Visually-aware Recommendation with Aesthetic Features. *CoRR* abs/1905.02009 (2019).

[30] Xuzheng Yu, Tian Gan, Yinwei Wei, Zhiyong Cheng, and Liqiang Nie. 2020. Personalized Item Recommendation for Second-hand Trading Platform. In *ACM Multimedia*. ACM, 3478–3486.