Learning Users' Preferred Visual Styles in an Image Marketplace

Raul Gomez Bruballa
rgomezbruballa@shutterstock.com
Shutterstock
Dublin, Ireland

Lauren Burnham-King
lburnhamking@shutterstock.com
Shutterstock
Dublin, Ireland

Alessandra Sala
asala@shutterstock.com
Shutterstock
Dublin, Ireland

ABSTRACT

Providing meaningful recommendations in a content marketplace is challenging due to the fact that users are not the final content consumers. Instead, most users are creatives whose interests, linked to the projects they work on, change rapidly and abruptly. To address the challenging task of recommending images to content creators, we design a RecSys that learns visual styles preferences transversal to the semantics of the projects users work on. We analyze the challenges of the task compared to content-based recommendations driven by semantics, propose an evaluation setup, and explain its applications in a global image marketplace.

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Computing methodologies** → *Machine learning*.

ACM Reference Format:
Raul Gomez Bruballa, Lauren Burnham-King, and Alessandra Sala. 2022. Learning Users' Preferred Visual Styles in an Image Marketplace. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3523227.3547382

1 RECOMMENDATIONS FOR CREATIVES

An image marketplace is a huge catalog of very diverse content in terms of semantics, where one can find from images of a train to images of a surgery, and also in terms of styles, where a train image can be a grayscale illustration or a landscape wide-angle image of a train crossing the Alps. Platform users are creatives whose content interests change rapidly since are dependent on the different projects they work on, and to navigate the catalog they use a search engine which provides content relevant to a given search query, ranked mainly by popularity.

As the catalog is diverse in terms of visual styles, creatives also are in terms of visual style preferences, and we hypothesise those are stable across projects. As an example, if a creative has licensed mostly illustrations in the marketplace, when he searches for "train" he will be probably interested in train illustrations, while if another user has licensed many wide-angle images of landscapes from Switzerland, when he searches for "train" he will probably be more interested in the train image crossing the Alps. In the proposed recommendations pipeline, we maintain the ability of creatives to search for content relevant for their projects, but we personalize search engine results to match each user's preferred visual styles, which we learn from their historical activity in the marketplace, aiming to serve content more relevant to them.

A key aspect of the presented `Visual Styles RecSys` is therefore to define image features encoding styles which are stable across creatives' projects. Features of different natures were found relevant and later tested experimentally. Among them general features as the image predominant colors or author country, features handcrafted from selected keywords as the image angle, or semantic verticals such as the image category. Note that, for an style feature to be useful for recommendations, it doesn't need to be stable across projects for all users: the model can learn which features or features interactions are useful to recommend relevant content to each user. However, it is critical to avoid features encoding project semantics, because they are too discriminative compared to style features and would prohibit learning users' preferred styles. As an example, if we include image keywords as a feature, the model would rapidly fit the keywords relevant for the projects a user worked on during the training window, instead of learning styles relevant for further projects.

2 VISUAL STYLES RECSYS

The pipeline to personalize content discovery at the image marketplace is depicted in Figure 1. It is a two-stages pipeline: First a search engine generates candidates relevant to a user search query; Then, the RecSys re-ranks those candidates, so that the ones inline with the user visual styles preferences are served first. The interest of maintaining the search engine in the recommendation pipeline is that creatives can still benefit from custom recommendations while they rapidly shift between projects. Additionally, they can utilize at the same time any search engine functionality, such as search filters or trending images boosting. It is indeed crucial that the search engine provides candidates with enough styles diversity so that they fit the preferred styles of any user.

`Visual Styles RecSys` has a two-tower architecture and learns users and images representations in a joint embedding space (Fig. 2). The user encoder is essentially a Multi Layer Perception: First, representations of each one of the user features of experimentally found dimensionalities are learned, then those representations are concatenated and go through the MLP until they are projected to the embedding space. In the image tower, representations of each one of the image features are learned, concatenated to be processed in parallel by a MLP and a Cross Network [2] with a single Cross Layer. Finally features coming out from the two branches are concatenated and projected into the embedding space. Cross Layers [2] explicitly model feature interactions, which is critical for our RecSys aiming to learn users' preferred styles based on interactions among image style features.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
RecSys '22, September 18–23, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9278-5/22/09.
https://doi.org/10.1145/3523227.3547382

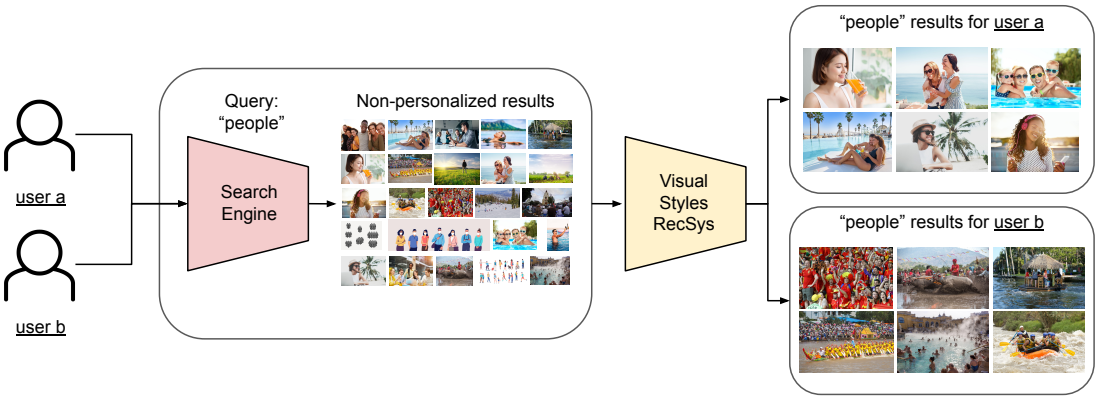

Figure 1: Recommendations pipeline. A user queries the search engine, which returns non-personalized results matching the query semantics (in the example, "people"). Then, Visual Styles RecSys re-ranks those results, and the ones inline with the user preferred visual styles are shown first. Note how in the example, while the search engine provides very diverse results, user a gets more stocky portraits with blurred backgrounds, while user b gets served first top-view cadid images of crowds.

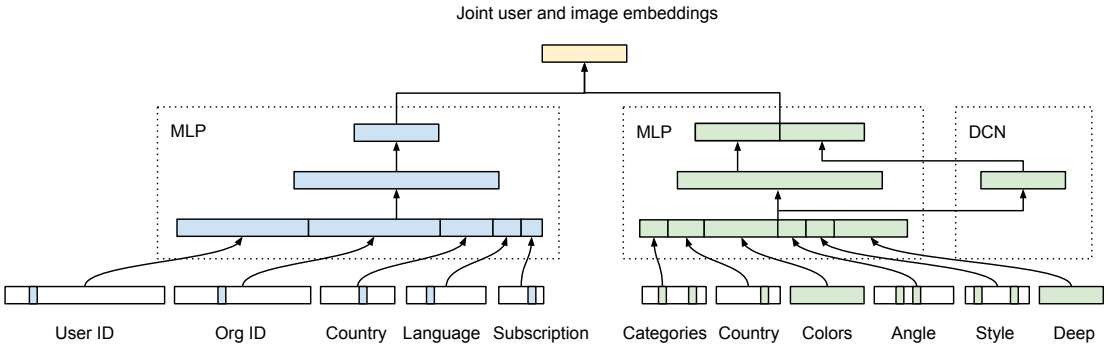

Figure 2: `Visual Styles RecSys` architecture. In blue, layers of the user encoder and in green, layers the image encoder. MLP (Multi Layer Perception) layers are linear layers with ReLU activations, and DCN (Deep Cross Network) layers are Cross layers.

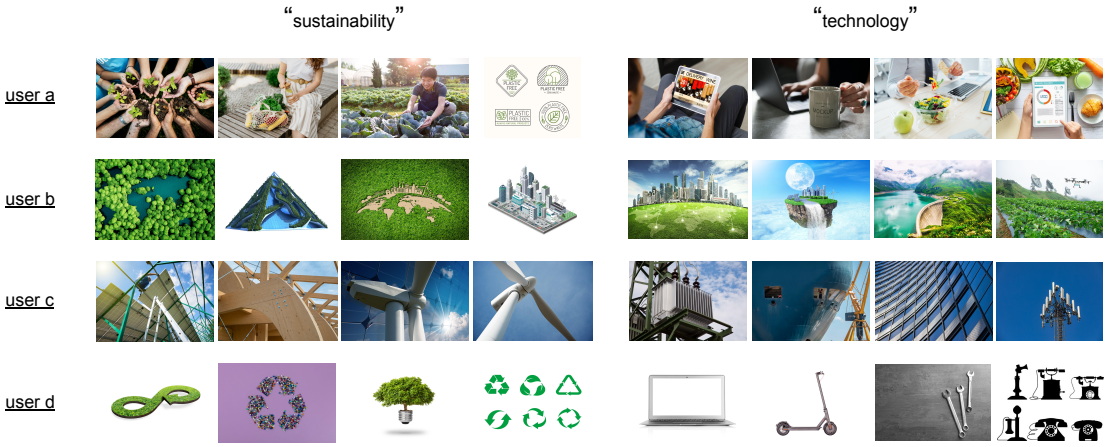

Figure 3: Top ranked images for four different users querying for "sustainability" and "technology" images. To better illustrate the learned styles, to get this qualitative results the search engine has been substituted by a simple keyword filtering: For each query, the test set is filtered by the querying keyword, and the images containing that keyword are re-ranked by the RecSys.

Table 1: Metrics leveraging different image features and a baseline where all users get recommended the most popular images.

Model	Acc@10	Acc@100	Catalog Coverage	ECS@10	Visual Diversity
`Popular Baseline`	0.02	0.19	2e-6	9.4	**0.413**
`VS RecSys: Deep`	0.093	0.655	35.75	55748	0.278
`VS RecSys: All but Deep`	0.021	0.250	14.38	16798	0.371
`VS RecSys: All features`	**0.144**	**0.890**	**51.72**	**85669**	0.309

3 EXPERIMENTS

As our experimentation data we collect a dataset consisting of $10M$ user clicks in images of $90K$ users during a 6 months time window. We split those in a training set and a testing set.

3.1 Evaluation Metrics

`Accuracy@k`: Given a test set with users and images they have clicked on, recommendations for those users are predicted. Then it is measured if the clicked image is within the top-k recommendations.
`Catalog Coverage`: Measures the percentage of test set images appearing in the test set users top-10 recommendations, monitoring the diversity of the recommendations among users.
`Effective Catalog Size (ECS)`: Proposed by Netflix [1], measures how diverse is the content shown to users at each k, being k the position in the recommendation ranking.
`Visual Diversity`: Average L2 distance within the deep features of top-10 recommendations for each user. Inspired in LPIPS [3], measures how diverse recommendations are for each user in terms of deep features.

3.2 Image Features Study

Table 1 shows the influence of different image features in the performance metrics. The model basing recommendations only on `Deep` features gets a lower score in `Visual Diversity` than the popular baseline. That's expected since any ranker leveraging deep features will tend to score lower than a popular ranking which does not. We can consider the `Visual Diversity` score of this model the lower bound, since it bases recommendations entirely on deep features, and any model successfully leveraging other features in addition to `Deep` features should score higher than this one. The model leveraging all the presented image features but `Deep` gets higher `Visual Diversity`. The reason is that `Deep` features are still correlated with the shallow features this model bases recommendations on. The `Accuracy` scores and the general diversity scores (`Catalog Coverage` and `ECS`) are lower than for the model based on `Deep` but still significant. That leads to two conclusions: First that this set of (shallow) features is relevant for recommendations and second that `Deep` features are key. However, as discussed before, exploiting `Deep` features rich representations comes with the risk of biasing recommendations towards the semantics of the projects users have worked on during the training window. `Visual Diversity` is excellent to monitor `Deep` features influence.

The model trained with all features scores in `Visual Diversity` significantly higher than the one leveraging solely `Deep` features and lower than the one not leveraging them. The scores show that it bases recommendations on the combination of `Deep` features and the rest of the presented ones, as pursued. It significantly outperforms the rest of the models in `Accuracy`, `Catalog Coverage` and `ECS`. That means that it successfully leverages features of diverse natures to provide recommendations that are of interest to users and diverse among them compared to the other models. Importantly, this means that leveraging all the presented features `Visual Styles RecSys` has successfully learnt users' preferred visual styles transversal to the projects they work on and, also, that the proposed evaluation strategy is useful to prove that by benchmarking models leveraging different features.

3.3 Qualitative Results

Figure 3 shows qualitative results for different users in the image marketplace demonstrating the diverse styles `Visual Styles RecSys` can learn. User a shows preference towards stocky close-up images featuring people and the gastronomy semantic vertical. Meanwhile, user b prefers top-view and wide angle images with greenish colors and digital art. User c is more interested in structures close-ups with bluish colors, while user d in signs and symbols and isolated objects with homogeneous backgrounds. Note that `Visual Styles RecSys` recommends images based not only on preferred users image features, but on complex styles learned based on the interactions of those features. Additionally, the model does not limit to learn a single preferred visual style for each user, but learns multiple preferred styles and, as a result, recommendations are diverse not only among users but also within them. As examples, in addition to the aforementioned preferences, user b has a lighter interest in isolated objects, and user a in illustrations.

PRESENTER CV

Raul Gomez Bruballa graduated in telecommunications engineering at UPC and coursed a master at the Computer Vision Center, where he also did his PhD together with Eurecat under an industrial PhD program. He has a strong research record on diverse tasks involving visual and textual data, such as self-supervised learning, multi-modal retrieval, or scene text detection. Currently he works at Shutterstock developing Recommender Systems.

REFERENCES

[1] Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms and Business Value and and Innovation. *ACM Transactions on Management Information Systems* (2016).
[2] Ruoxi Wang, Rakesh Shivanna, Derek Z. Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. *ACM World Wide Web Conference* (2021).
[3] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Conference on Computer Vision and Pattern Recognition* (2018).