Vis2Rec: A Large-Scale Visual Dataset for Visit Recommendation

Michaël Soumm[1] michael.soumm@cea.fr
Adrian Popescu[1] adrian.popescu@cea.fr
Bertrand Delezoide[2] bertrand.delezoide@amanda.com

[2]Amanda, 34 Avenue Des Champs Elysées, F-75008, Paris, France
[1]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

Abstract

Most recommendation datasets for tourism are restricted to one world region and rely on explicit data such as check-ins. However, in reality, tourists visit various places worldwide and document their trips primarily through photos. These images contain a wealth of raw information that can be used to capture users' preferences and recommend personalized content. Visual content was already used in past works, but no large-scale publicly-available dataset that gives access to users' personal images exists for recommender systems. As such a resource would open-up possibilities for new image-based recommendation algorithms, we introduce Vis2Rec, a new dataset based on visit data extracted from users' Flickr photographic streams, which includes over 7 million photos, 36k recognizable points of interest, and 14k user profiles. Google Landmarks v2 is used as an auxiliary dataset to identify points of interest in users' photos, using a state-of-the-art image-matching deep architecture. Image-based user profiles are then constituted by aggregating the points of interest detected for each user. In addition, ground truth visits were determined for the test subset in order to enable accurate evaluation. Finally, we benchmark Vis2Rec using various existing recommender systems, and discuss the possibilities opened up by the availability of user images, as well as the societal issues that come with them. Following good practice in dataset sharing, Vis2Rec is created using only freely distributable content, and additional anonymization is performed to ensure the privacy of users. The raw dataset and the preprocessed user profiles will be publicly available at `https://github.com/MSoumm/Vis2Rec`.

1. Introduction

Points of interest (POIs) are a central part of tourist experiences. Ideally, tourists should receive personalized recommendation in order to discover new places which are most interesting to them. Such personalization can be achieved by leveraging user profiles that encode their tourist preferences [10, 53]. Mainstream recommendation methods relied on a form of matrix factorization to propose personalized content [21, 14, 45], while more recent methods use deep learning algorithms to improve the performances of recommender systems [9, 44], and their effectiveness is largely determined by the quality and richness of the available profiles. While progress was made for the profile creation step, again through the introduction of deep learning techniques [5], this component needs further exploration. In this work, we propose to examine if user profiles constructed by identifying POIs on users' photos are suitable for POI recommendation. The main contribution of this work is the introduction of Vis2Rec, a new photo visual dataset designed to study this hypothesis. We also hope that Vis2Rec will stimulate research towards the creation of user-based recommender algorithms.

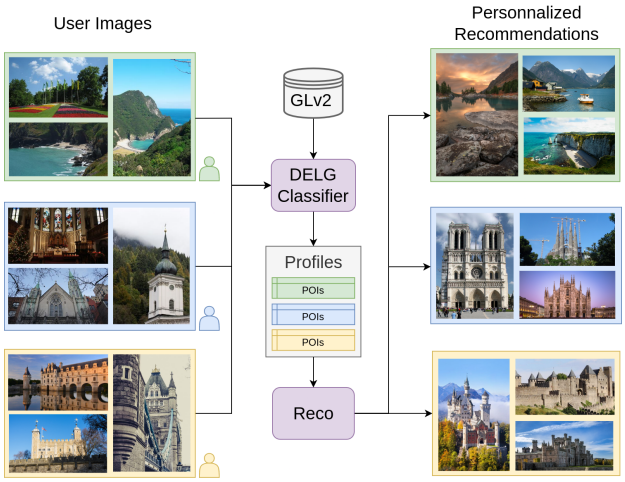

Figure 1: Proposed use-case of Vis2Rec .

A simplified illustration of Vis2Recproposed usage is proposed in Figure 1. Collected from Flickr, it includes 7,158,454 total photos, 14,600 visiting users, 36,111 POIs, and 421,065 unique POI visits. Data collection was restricted to Creative-Commons-licensed content in order to enable its public sharing.

Secondly, we examine the role of visual mining in the profile constitution process. Following recent trends [3], we use a deep learning model which recognizes 81k POIs by leveraging the Google Landmarks Dataset v2 [54]. Because a visual classifier transforms raw data into profiles without requiring effort from the users, such a method is suitable for large-scale real-world applications. As a result, the ability of the model to recognize many POIs is the driving factor in creating detailed profiles which cover a large number of cities or towns (referred to as cities hereafter).

The proposed profile extraction process is instantiated for the creation of Vis2Rec. However, it is more generally applicable to create rich tourist profiles with minimal effort on the user side. This is because raw data are transformed into actionable cues with no effort required on the user side, as opposed to explicit contributions required by check-ins based datasets such as Gowalla or Foursquare [57].

We compare 8 existing recommendation methods that provide promising performances according to existing benchmarks[1] in section 3. Profiles are solely based on the visual identification of POIs on the user images. The results show that information extracted from the visual data is suitable for the recommendation task, and that Vis2Rec provides a new benchmark for state-of-the-art recommender systems. To facilitate reproducibility and stimulate future research, the dataset will be made publicly available at `https://github.com/MSoumm/Vis2Rec`.

2. Related Work

Existing POI datasets. POIs are an important component of tourist visits and their recommendation has received strong attention from the research community [10, 53]. A recent review discusses POI recommendation based on multimedia content [10], underlining the central role of visual content in recording tourist visits. As a result, it is natural to leverage visual datasets in order to elicit user preferences in the direction of personalizing their tourist experiences. In an earlier work (2009), the authors of [19] collected 400,000 images from tourism blogs. Photo content was mined and integrated into a graph-based framework to propose personalized visits in a few dozen large cities. A large-scale dataset was collected from Panoramio before 2010 and leveraged for POI recommendation named Photo2Trip [28]. This dataset included over 20 million geotagged images and 30,000 POIs which covered over 100 countries. POI discovery was done based on a clustering of geotags associated to images, and not on an analysis of the image content. A more refined recommender based on geotags was proposed in [2], where the authors introduce a semantic component in user profiles. An important hypothesis made in [2, 28] is that geotags are available for all photos. However, this is often not the case either because photos are taken with devices that do not record geolocation, or because users are reluctant to share both types of data [46]. An Instagram dataset introduced in [51] contains images taken in two cities (New York and Chicago), and a visual model pretrained on ImageNet LSVRC [39] is used to describe image content. While the authors stated that the dataset will be publicly available, this is no longer the case, probably due to copyright and user consent issues that are related to the choice of the data source.

The unavailability of large-scale POI-related datasets makes the comparison of methods and reproducibility of results difficult. We introduce such a dataset to facilitate the sound development of future works. Our work is informed by challenges that prevented the publications or led to the withdrawal of past datasets. The measures implemented to respond to these challenges pertain to data provenance, rights, and processing, and are detailed in subsection 3.4.

Also of interest are datasets such as Gowalla checkings [6] and Foursquare Complete [56]. They are mined from location-based services and are based on user checkins. While interesting, there are two important differences with our dataset: (1) they require an explicit contribution from users for checking-in, and (2) their focus is not on tourism, but rather on commercial activities (Gowalla) or on cultural mapping by local users (Foursquare). Moreover, due to their size and nature, the heavy preprocessing needed to convert raw data into POI visits leads to often considering only a localized slice of these datasets [57].

POI mining. This is a central component of our approach. A comparison of data sources used for POI recommendation [40] underlines their complementarity. The authors of [24] introduced an interesting approach that combines textual, visual, and user data to associate Instagram photos to POIs. Then, the obtained information was aggregated in user profiles which express their preferences. One hypothesis made is that textual data are available, which is true for a part of the images shared on social network, but not for the majority of user images, which are simply stored.

Visual POI recognition is an appealing alternative because it only requires the photos themselves, and no effort from the user side to build the profile. The main challenge here is to ensure that the recognition process is sufficiently accurate for a search space which includes a very large number of POIs. The availability of rich datasets such as Google Landmarks [54] facilitated the development of deep learning approaches to tackle POI recognition at scale. The task can be implemented using either classification or matching approaches. Recent classification approaches [8, 55] use deep architectures such as EfficientNet [47] or visual transformers [27], in isolation or ensembled, to automatically label POIs in images. This process is fast since it only requires an inference for test images. However, it re-

[1] `https://paperswithcode.com/task/recommendation-systems`

Dataset name	Domain	#Users	#Items	#Interactions	Additional information
Amazon clothing [36]	E-shopping	58,197	44,310	422,474	Item images, Item features
MovieLens20M [11]	Movie	20,720	136,677	20M	Item features
Netflix [1]	Movie	463,435	17,769	100M	Item features
Foursquare (POI) [57]	Travel	2,321	5,596	194,108	User relations
Gowalla (POI) [57]	Travel	10,162	24,250	456,988	User relations
Vis2Rec (ours)	Travel	14,600	36,111	421,065	User images, Item images, Item features

Table 1: Comparison of Vis2Rec (filtered for recommendation) with a few existing recommender system datasets

quires a relatively large number of labeled images per class for a reliable classification, and this condition is not met for many POIs. Deep visual matching [32] is mainly based on local content descriptors which are matched in a pair of images. The challenge here is to make the matching process scalable since each image is potentially compared to all available reference images. Recent methods [3] reduce the complexity of the process by implementing a two-stage matching process: a lightweight global description is first used to preselect similar reference images. Local features are then matched only for the most promising preselected candidates. Such methods are more adapted here because they can be run even for scarcely represented POIs and can be deployed to POIs which were not explicitly learned, contrary to classification methods.

Recommendation. The increase of available user data and processing power in the last decades have led recommender systems to mostly rely on Collaborative Filtering (CF) techniques. While neighborhood approaches such as user-knn remain simple yet effective baselines, Matrix Factorization (MF) methods [21, 17, 14, 23] have been predominant since the Netflix Prize challenge [1]. More recently, new recommender systems leverage advances in deep learning by using VAE architectures [25, 26, 49, 44] or graph structures [13]. While most of these methods primarily rely on the user-item matrix, additional visual data can be used to enhance recommendation [36, 12, 33]. In these works, only the item visual features are used. However, authors of [52], whose setup is the closest to our work, show that using user images is useful for the POI recommendation task, but unfortunately, the used dataset cannot be found anymore. The incentive to use mostly item features as additional information mainly comes from the available data. A comparison of the most popular recommendation datasets (Table 1) shows that the available information usually includes a large-scale user-item matrix, along with item features and/or images, and at best user graph relations. Despite the setup of [52] being promising, the lack of user image data is an obstacle to future research in that direction.

3. Vis2Rec Dataset

The goal of Vis2Rec is to provide a realistic and sustainable testbed for visit recommendation, based on user images. To meet this objective, we need to address technical, legal, and ethical challenges.

The dataset is built to propose recommendations at scale, and after the correct processing for recommendation, it caters to at least 36,111 POIs in 5,012 cities. These POIs are taken from Google Landmarks v2 (GLv2) [54] in order to enable large-scale visual POI recognition. The size of the user set is also important in order to capture diversified user preferences. Preprocessed Vis2Rec includes a total of 14,600 users, 829,673 POI-associated user images, and over 6M additional images.

Sustainability is ensured by implementing a legally-compliant data collection and distribution process. The dataset includes only distributable images which were taken on visit days. Equally important, face de-identification was applied to ensure the anonymity of the users.

We describe the main steps of the dataset constitution and packaging below, and the data collection and structuring pipeline is summarized in Figure 2.

3.1. Initial data collection

POI set. GLv2 [54] is one of the largest publicly available POI-related dataset, which was collected from Wikimedia commons. We use the "clean" subset which includes a total of 1,580,470 images which represent 81,313 POIs. GLv2 is therefore adapted for the creation of a comprehensive visit recommendation dataset, such as Vis2Rec. To perform efficient data queries, we need to enrich this dataset by mining information from the Wikipedia pages associated to POIs. The resulting dataset includes the name of the POI (with translations, when available), its associated GPS coordinates, and the closest city from the Geonames[2] list of 139,439 cities which have at least 1000 inhabitants.

User set. Flickr offers an easy-to-use API for a large collection of images and associated metadata, and is as such a very adapted data source to our work setup. We launch Flickr API queries with the POI name(s), using a 3 km radius around the coordinates. Queries are limited to photos distributed under Creative Commons licenses to ensure that they are redistributable. Metadata for up to 5000 photos is collected for each POI, containing photo ID, user ID, and user tags, as well as geographic coordinates of the photos. This process provides an initial list of 20k preselected users.

[2] https://www.geonames.org/

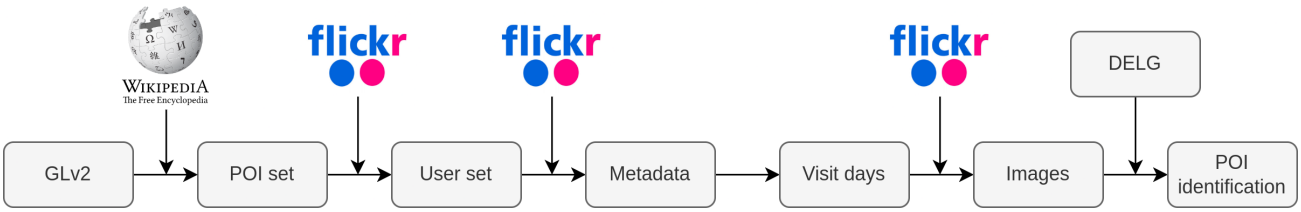

Figure 2: Data collection and annotation pipeline.

3.2. Domain-related data selection

The image collection should be focused on tourist visits. More specifically, we collect all the photos corresponding to a potential visit day, determined by generating coarse POI predictions for each image. A day is kept if it includes at least one POI name in the image tags. Since POI names are often ambiguous [35, 42], further post-processing is needed to disambiguate potential POI matches. Whenever geolocation is available for at least one photo taken during one day, it is used to check for POIs which are located within a radius of 10 kilometers. If geolocation is not available, we resort to text-based matching which uses a probabilistic geographic language model [42]. This model associates the visit day with a list of probable cities based on the aggregation of the location probabilities of the tags used during a tested day. A geolocated subset of metadata is used to determine a threshold which provides a good balance between precision and recall for detected visit days.

This matching provides a text-based profile of each user [22] which is used to select interesting users for the visual dataset. The direct use of text-based profiles for recommendation [22, 34] is possible but is suboptimal since users are required to provide explicit textual annotations of their visits, which often leads to incomplete profiles. The resulting intermediate dataset includes 17k user profiles and a total of 27k text-annotated POIs.

3.3. Visual matching of POIs

Vis2Rec is intended for recommendation based on the sole use of photo content and we should make no assumption regarding the availability of textual annotations or geolocation for the dataset. This is important in practice in order to design a profiling pipeline that does not require any effort from the users. Consequently, we collect images for the visit days identified in the intermediate dataset based on tags (Subsection 3.2). These photos are then compared to POI images from Google Landmarks v2 dataset [54] using a DELG descriptor [3].

Visual matching procedure. Visual matching is performed using DELG [3], which achieves state-of-the-art single model instance-level recognition on GLv2. We use the model only for inference since the pretrained weights on GLv2 can be found in the official implementation [3]. The visual matching of candidate and reference images is done in two steps:

1. a 2048-dimensional global embedding is used to select a subset of similar reference images from GLv2 for each candidate image in which POI occurrences are searched. Following common practice, the top-20 most similar reference images are retained for the second step.
2. a geometric verification process based on 128-dimensional local descriptors provided by DELG is performed to refine the list of similar reference images. The final ranking is based on the number of matched keypoints between the candidate and the reference images.

This two-step process is needed since global retrieval is fast but potentially prone to errors, while geometric verification is slow but accurate. Each candidate image is paired with the reference image that has the highest matching score, and attributed with the POI represented by this reference image. The number of keypoints can be used as a confidence estimator for the quality of visual matching.

Results. Since DELG was pretrained on the same POI set as Vis2Rec , the visual matching procedure has good qualitative results (see Figure 3). Correct identification is possible for a wide range of setups, including outdoor landscapes, indoor architectures, as well as difficult lighting conditions. However, this process is far from perfect and fails in particular situations (Figure 3). By analyzing the results of the visual matching, we can identify three types of recurring errors: (1) objects which occur in different regions of the world and are representative for POIs (Figure 3 (d)); (2) visually similar objects which are specific to a city (Figure 3 (e)); and (3) visually similar POIs (Figure 3 (f)).

The first type of error can be reduced by removing GLv2 reference images which match target images located in different parts of the world. To do this, we use a geolocated validation set and remove any reference image which was matched only to POIs farther than 15km away at least 5 times. The remaining spatial aberrations are removed by selecting the most confident POI detection for each day and removing detections corresponding to POIs farther than 100km from it. This geographic filtering removes over 1 million images.

[3] https://github.com/tensorflow/models/tree/master/research/delf

Figure 3: Examples of visual matches provided by DELG. The model recognizes correctly: (a) outdoor landscapes; (b) indoor scenes; (c) different lighting conditions. Errors can be caused by:(d) same objects in different places; (e) Signs with identical features; (f) similar architectures.

The second type of error is the most difficult to handle since neither a spatial criterion nor a good matching score threshold can deal with them.

The third type of error is usually associated to lower matching scores. By manually verifying a few hundred matched image pairs, we observe that a matching score of 30 leads to an accuracy of at least 98%. Interestingly, this coincides with the threshold chosen in the GLv2 article [54] to generate the "clean" subset, and to the threshold that leads to the best recommendation results (see Section 4). In the rest of this work, this is the default chosen threshold.

3.4. Data distribution

We present dataset-level statistics for detected POIs and user visits. These statistics are obtained after applying the visual matching error mitigation measures described in Subsection 3.3, and lead to a dataset comprised of 36,111 unique POIs, depicted on 820,593 images, corresponding to 421,065 unique user visits. Since these statistics highly depend on the chosen matching threshold, the distributed dataset contains all of the POI predictions without any filtering to allow for further research and POI discovery.

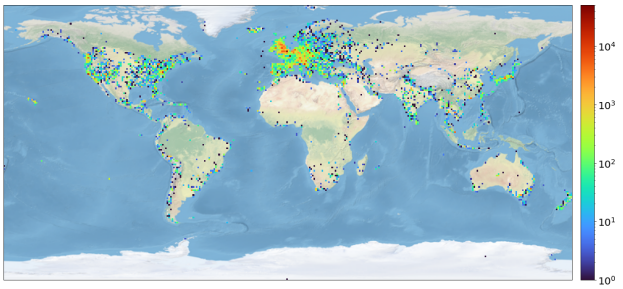

Figure 4: Spatial distribution of identified visits. Darker points correspond to a large number of identifications.

Identified POIs. Figure 4 illustrates the distribution of identified visits across the world, along with the associated number of detections. The obtained distribution is in line with global tourist visit trends [50], and shows a strong concentration of POIs in Western Europe, East and West coasts of North America, and Eastern and South-Eastern Asia. The distribution is also influenced by Flickr usage trends, and confirms previous analyses of geolocated photos shared on this platform [7, 35]. The distributions of the number of identified POIs and the number of visits per city are proposed in Figures 5 and 6, respectively.

Both of them exhibit long-tail shapes, with a large number identified POIs and of visits concentrated in large tourist hotspots, such as London, Paris, New York City, and significantly fewer visits associated to the other cities. More details about the visited POIS and visits in the different cities are provided in the supplementary material.

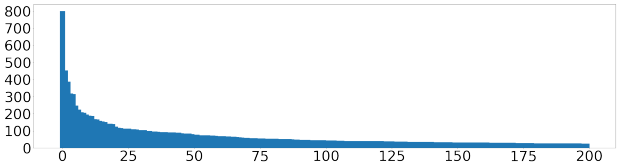

Figure 5: Distribution of the number of identified POIs in the top 200 cities.

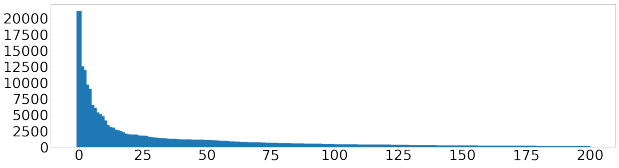

Figure 6: Distribution of the number of user visits in the top 200 cities

User visits. User profiles generated in Vis2Rec are rich and diversified. First of all, 84% of the users visited at least 5 POIs, a threshold commonly used in recommender systems for filtering purposes, while the median user visited 16 distinct POIs. Secondly, 95% of the users visited more than one city, 8 being the number of cities visited by the median user, resulting in rich travel profiles. These observations can be easily explained by the fact that travel images are often uploaded to Flickr to highlight their extraordinary nature. Therefore, one should keep in mind that Vis2Rec does not contain images that are representative of the everyday life of its users, but more of their vacation travels.

Additional images. Confident POI detections account for 11% of the 7,158,454 total images. We estimate that between 1 and 2 million other images could depict POIs, and counting them as valid by lowering the matching score threshold would increase the POI set to around 60k unique POIs. However, this introduces many false positives in the user profiles, resulting in lower recommendation performances. As per this observation, a threshold of 30 matching keypoints is kept throughout our work. The remaining images are non-POI personal user photos and are distributed for potential further work.

3.5. Dataset partition and annotation

Splits. The dataset is split into train, validation, and test subsets to enable the application of learning-based recommendation methods to it. The validation and test subsets are further divided into *inputs* and *targets*, the former being used as the user known profile to get recommendations, and the latter to calculate metrics and benchmark recommender systems. The *target* set is verified to ensure the meaningfulness of the recommendation benchmarks. We preselect the top 200 cities based on their number of POIs and look for users that exhibit multiple visits in these cities. We isolate 2,100 such users and determine their ground truth visits either automatically or with manual annotation.

Automatic annotation. In preliminary experiments, we analyzed random samples of target-reference image pairs provided by the geometric matching process. We partitioned the matched pairs into bins based on their matching score, each bin corresponding to a 10-keypoints window. We then drew 500 random samples from each bin and performed a manual verification of the matched pairs. The results showed that the visual matching has an accuracy of over 99% when the number of matched keypoints is larger than 40. As the test set requires more confident annotations than the train set, we decided to label all pairs which share more than 40 keypoints as correct.

Manual annotation. A manual annotation process is run for the remainder of the target subset of validation and test users. A total of 10k image pairs with a matching score lower than 40 are manually verified. The task is relatively simple since annotators need to decide whether the two images of a target-reference pair depict the same POI or not. Three annotators verify each pair and we consider the match to be correct if at least two of them label it correct. More details about the annotation process and interface are provided in the supplementary material.

Split	#Users	#Items	#Interactions	#Images
train	13,066	34,291	343,286	5,914,005
test/val	1,534	16,822	54,743	951,012

Table 2: Splits of Vis2Rec processed for recommendation.

As a result of this filtering, only 1,534 users with more than 5 annotated POIs remain, as its a common threshold for recommender systems [20]. The users' annotated POIs constitute their *target* set and account for 30% of their visits, the remaining 70% being used as the *input* set.

The resulting splits are described in table 2. For our experiments, we use a test set size of 1000 and use the 534 remaining users for validation.

3.6. Dataset compliance

First, Vis2Rec was collected via the official Flickr API, a data source which allows the constitution of datasets made of data originally shared by its users. For instance, the well-known YFCC100M dataset [48] was also collected from Flickr and is still available today. Second, we keep only images which are shared under Creative Commons (CC) licenses in order to enable lawful redistribution of content. The dataset will be published using a license that is compatible with the most restrictive CC licenses included in Vis2Rec, and commercial reuse will be notably not permitted. Third, we will enforce the data minimization principle defined in Article 5 of the General Data Protection Regulation[4], and share only the data needed for the POI recommendation task. The dataset includes only images taken on days that correspond to tourist visits. A qualitative exploration of Vis2Rec showed that it contains many personal images. As such faces will be de-identified [29] in the dataset to protect the anonymity of the depicted persons.

4. POI recommendation

4.1. Tasks and Metrics

Data used. As described in section 3.4, the POIs identified in Vis2Rec by DELG create a subset suitable for recommender systems, but also includes other images which could further enrich profiles. Since the main objective of the paper is to describe the dataset, we benchmark recommendation systems that work on top of POI-based profiles. Usage of the additional data is left for investigation in future work. Unfortunately, since there are no other available

[4] https://gdpr-info.eu/art-5-gdpr/

POI image recommendation datasets, a comparison of this benchmark with other data sources would not be fair and thus is not performed here.

Notations. We denote $\mathcal{U} = \{u_1, ..., u_n\}$ and $\mathcal{I} = \{v_1, ..., v_m\}$ the sets of users and items (POIs) respectively. We make an assumption of implicit positive feedback since we equate photo uploads with an interest for the visited items. Also, since no range of user experience can be determined, our data is binary, contained in a user-item matrix $R = [r_{ij}]$. A user u_i is therefore encoded by a sparse binary vector representing their inferred visits. We will denote by S_{u_i} the set of items we want to predict for the user u_i.

Recommendation pipeline. Algorithms are trained using the *train* set visual predictions. An item from the training set is considered visited when its visual matching score is above 30. Then, during the test phase, algorithms receive a test input sparse vector u_i^{input} representing the known visits for user u_i, and predict ratings $\hat{r}_{ij}$ for each $v_j \in \mathcal{I}$. Item are then ordered by decreasing rating score to calculate ranking metrics, namely *Recall@N* and *NDGC@N*, commonly in recommender system benchmarking [41]. To obtain a fair comparison to baseline methods, we filter the predicted items to the POIs located in the cities visited by the test users. More detailed discussion of metrics can be found in the supplementary material.

4.2. Recommendation methods

We aim to benchmark Vis2Rec for recommendation systems, based on an interaction matrix which associates users and POIs. We discuss the tested methods below.

Oracle. A perfect recommender system that produces a list of train items truly visited by the user. Since some test items may not be present in the training set, and ranking metrics use a top-N recommendation list (which, depending on N, may be too small to cover all target items), the Oracle system provides us with a performance upper bound.

MostPop. The "travel guide" recommendations, which recommends the most popular items for everyone. Popularity is computed based on the count of training users who visited the POI. This is a strong baseline since it aggregates the interests of a large community of travelers [34].

User-KNN. Computes the similarity (usually cosine) $Sim(u_i, u_{i'})$ between the vector of known items of u_i and the vectors of all the train users $u_{i'}$, then predicts the score

$$\hat{r}_{ij} = \frac{\sum_{v_{i'} \in \mathcal{N}(u_i)} Sim(u_i, u_{i'}) \times r_{i'j}}{\sum_{u_{i'} \in \mathcal{N}(u_i)} |Sim(u_i, u_{i'})|}$$

where $\mathcal{N}(u_i)$ is the top-k neighborhood of u_i.

MF[21]. A method popularized by the Netflix challenge. Learns by gradient descent latent vectors $\mathbf{u}_i$ and $\mathbf{v}_j$ for users and items to optimize :

$$\hat{r}_{ij} = \mathbf{u}_i^T \mathbf{v}_j + \frac{\lambda}{2}(||\mathbf{u}_i|| + ||\mathbf{v}_j||)$$

WMF[17]. In the case of implicit feedback, observed interactions contain more information than unobserved ones. Weighted Matrix Factorization weighs the MF loss so as to penalize observed interactions more than unobserved ones.

BPR[38]. Bayesian Personalized Ranking turns implicit feedback into ordinal preferences. It considers a user u_i who interacted with item v_j and did not interact with item v_l to maximize the joint likelihood over all triplets:

$$\prod_{(i,j,l)} \sigma(\mathbf{u}_i^T \mathbf{v}_j - \mathbf{u}_i^T \mathbf{v}_l)$$

NeuMF[14]. Neural Matrix Factorization treats recommendation as a classification task with respect to a binary r_{ij}. It combines shallow (Generalized Matrix Factorization) and deep (MLP) user and item representations and is known to outperform MF in an implicit rating setup, such as ours.

EASE[45]. A popular and simple recommender system with close to state-of-the-art results on many recommendation datasets. It computes an item-item weight-matrix, similar to SLIM [31], but much more efficiently, by solving:

$$\min_B ||R - BR||_F^2 + \lambda ||B||_F^2, \text{ s.t. } \text{diag}(B) \qquad = 0$$

with a closed-form solution.

RecVAE[44]. Variational Auto-Encoders are deep learning architectures that learn a latent space with a better structure than simple auto-encoders. By leveraging a sampling mechanism in the latent space, they are less prone to overfitting and achieve state-of-the-art results on MovieLens 20M and the Netflix datasets. We use the RecVAE variant, which is inspired by β-VAE [16] and denoising-VAE [18], and is adapted for an implicit binary-data setup.

4.3. Benchmarking

Methodology. We trained all methods described in section 4.2 using the Cornac[5] library. RecVAE, which was not present natively, was re-implemented. All methods were trained on the *train* set, and fine-tuned on the *validation* set when needed. The optimized hyperparameters sets found for each method are detailed in the supplementary material.

Results. Results for both benchmarking tasks on the *test* set are reported in table 3. For trainable methods, five randomized runs were aggregated and averaged performance is reported. The standard deviation for each metric is in the order of 0.005. The best performance according to all metrics is obtained with RecVAE, followed by EASE, with a significant gain associated to the first method. Since our data is binary and positive only, factorization methods underperform even compared to MostPop. The high scores of

[5] https://github.com/PreferredAI/cornac

	Recall@20	*Recall@50*	*NDGC@20*	*NDGC@50*
Oracle	0.9393	0.9675	0.9816	0.9794
MostPop	0.2777	0.4509	0.2240	0.2897
User-KNN	0.2745	0.4518	0.1956	0.2642
MF[21]	0.2196	0.4239	0.1255	0.2043
BPR[38]	0.2806	0.4636	0.2240	0.2939
WMF[17]	0.2735	0.4629	0.2009	0.2743
NeuMF[14]	0.2557	0.4279	0.2023	0.2665
EASE [45]	0.2979	0.4787	0.2475	0.3176
RecVAE [44]	**0.3410**	**0.5140**	**0.3003**	**0.3644**

Table 3: Performance of tested recommender systems on Vis2Rec .

			ΔRecall@20	*ΔRecall@50*	*ΔNDGC@20*	*ΔNDGC@50*
POI Ablation	25%	MostPop	0.00	0.00	0.00	.00
		EASE	-0.02	-0.03	-0.02	-0.02
		RecVAE	-0.02	-0.03	-0.04	-0.03
	50%	MostPop	0.00	0.00	0.00	0.00
		EASE	-0.04	-0.05	-0.05	-0.05
		RecVAE	-0.07	-0.01	-0.07	-0.07
User Ablation	25%	MostPop	0.00	0.00	0.00	0.00
		EASE	-0.01	-0.01	-0.01	-0.01
		RecVAE	0.00	0.00	0.00	0.00
	50%	MostPop	0.00	0.00	0.00	0.00
		EASE	-0.02	-0.02	-0.02	-0.03
		RecVAE	-0.02	-0.02	-0.02	-0.02

Table 4: Results difference when ablating the user profiles or the number of users. The relative difference with results from Table 3 is presented.

MostPop also indicate that users tend to deviate only moderately from the average behavior of visitors, which favors famous POIs over the rest of the cultural offer of the modeled cities. This makes the tasks more interesting since the information needed for better performances has to be extracted through more advanced methods. The user-item space modeling done with RecVAE meets this requirement since this method clearly outperforms MostPop. The results reported here constitute a sound baseline for future works which will use Vis2Recsince they cover a large panel of methods.

Ablation study. Table 4 describes the performance differences for recommendation when ablating POI identification and training profiles from Vis2Rec , respectively. We report results for the best two algorithms determined on the full data, along with those for MostPop baseline. MostPop is robust to the ablations, and this indicates that the dataset is large enough to create a stable popularity-based ranking of recommended POIs. In contrast, EASE and RecVAE are negatively impacted by ablations. This is intuitive since they rely on finer-grained cues learned from the user-item interactions. The ablation of 50% of identified POIs has the strongest impact, with a performance reduction for RecVAE of at most 7 percentage points. However, the corresponding reduction is only 3 percentage points when 25% of the POIs are removed. This observation, along with the stable results obtained when removing 25% user profiles shows that the total size of the training set allows for a robust benchmarking of the recommendation algorithms.

5. Ethics and Societal impact

Recommender systems provide a useful service to users, but their widespread use has also generated strong concerns due to the privacy-personalization trade-off that they require [37], and to biases that they generate [4]. Aware of the first challenge, we propose a dataset that includes only public content that is redistributable, we limit the distribution of images to those taken on visit days and anonymize all images which include faces, as described in Subsection 3.6. Moreover, the proposed use-case of Vis2Rec could be acheived in a scenario where profile construction and recommendation are made on the users' device.

The negative effect of biases generated by recommendation was notably highlighted for the political domain [30] or e-commerce-related over-consumption [15]. The latter risk can occur for the recommendation of tourist visits and can lead to an increased carbon footprint of the users, but can be mitigated by favoring the recommendation of visits to nearby destinations. Such proposals are in line with tourist trends observed following the COVID-19 pandemics [43] and are likely to be accepted by users.

6. Conclusion

We introduce Vis2Rec, a dataset for visit recommendation, to fill the gap generated by absence of a large-scale publicly-available resource in this domain. We describe its constitution, the measures implemented to ensure its sustainable distribution, an evaluation methodology, and a benchmark of a diverse set of recommendation algorithms. The obtained results show that the proposed task is challenging, and thus future research is needed to improve further improve performance.

Encouraged by the promising results reported here, we discuss potential improvement axes. First, we obtained recommendation results based on the identification of POIs in images by using a recent visual matching algorithm [3]. The distribution of images facilitates the inclusion of stronger algorithms that are likely to be developed in the future. Second, images contain additional cues which could be leveraged in order to obtain more comprehensive profiles [52], and ultimately improve the proposed recommendations. Such enrichment is made possible by the proposed distribution of all Flickr images uploaded by the users for each visit day. Third, Vis2Rec comes with additional data which could be leveraged. Simple use of geo-temporal data was proposed here, but more refined techniques, such as [23] would probably prove beneficial. Data at higher semantic and finer-grained semantic levels could also be extracted for Vis2Rec . Finally, we fed user profiles to a diversified range of recommendation algorithms to highlight the usage of the datasets. It would be interesting to test other additional recent algorithms, such as [26, 12, 36] and their future developments, to improve performance.

References

[1] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer, 2007.

[2] Guochen Cai, Kyungmi Lee, and Ickjai Lee. Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos. *Expert Systems with Applications*, 94:32–40, 2018.

[3] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020.

[4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*, 2020.

[5] Lei Chen, Lu Zhang, Shanshan Cao, Zhiang Wu, and Jie Cao. Personalized itinerary recommendation: Deep and collaborative learning with textual information. *Expert Systems with Applications*, 144:113070, 2020.

[6] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: Friendship and mobility: User movement in location-based social networksfriendship and mobility. In *User Movement in Location-Based Social Networks ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[7] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770, 2009.

[8] Shubin Dai. 2nd place solution to google landmark recognition competition 2021. *arXiv preprint arXiv:2110.02638*, 2021.

[9] Aminu Da'u and Naomie Salim. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53(4):2709–2748, 2020.

[10] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, 53(5):1–38, 2020.

[11] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[12] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[13] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation, 2020.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering, 2017.

[15] Morrant Hemans and Dickson Kofi Wiredu Ocansey. The impact of recommendation system overuse on the subjective wellbeing of internet users. 2021.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[17] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

[18] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework, 2015.

[19] Rongrong Ji, Xing Xie, Hongxun Yao, and Wei-Ying Ma. Mining city landmarks from blogs by graph modeling. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 105–114, 2009.

[20] Joseph A Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In *Proceedings of the international workshop on Reproducibility and replication in recommender systems evaluation*, pages 23–28, 2013.

[21] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[22] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. Travel route recommendation using geotagged photos. *Knowledge and information systems*, 37(1):37–60, 2013.

[23] Xutao Li, Gao Cong, Xiaoli li, Tuan-Anh Pham, and Shonali Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. *SIGIR*, 09 2015.

[24] Xutao Li, Tuan-Anh Nguyen Pham, Gao Cong, Quan Yuan, Xiao-Li Li, and Shonali Krishnaswamy. Where you instagram? associating your instagram photos with points of interest. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1231–1240, 2015.

[25] Xiaopeng Li and James She. Collaborative variational autoencoder for recommender systems. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 305–314, 2017.

[26] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698, 2018.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[28] Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 143–152, 2010.

[29] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. Cfa-net: Controllable face anonymization network with identity representation manipulation. *arXiv preprint arXiv:2105.11137*, 2021.

[30] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

[31] Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. pages 497–506, 12 2011.

[32] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[33] Chanyoung Park, Donghyun Kim, Jinoh Oh, and Hwanjo Yu. Do" also-viewed" products help user rating prediction? In *Proceedings of the 26th International Conference on World Wide Web*, pages 1113–1122, 2017.

[34] Adrian Popescu and Gregory Grefenstette. Mining social media to create personalized recommendations for tourist visits. In *Proceedings of the 2nd international conference on computing for geospatial research & applications*, pages 1–6, 2011.

[35] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93, 2008.

[36] Ruihong Qiu, Sen Wang, Zhi Chen, Hongzhi Yin, and Zi Huang. CausalRec: Causal inference for visual debiasing in visually-aware recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, oct 2021.

[37] Naren Ramakrishnan, Benjamin J Keller, Batul J Mirza, Ananth Y Grama, and George Karypis. When being weak is brave: Privacy in recommender systems. *arXiv preprint cs/0105028*, 2001.

[38] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[40] Maria Henar Salas-Olmedo, Borja Moya-Gómez, Juan Carlos García-Palomares, and Javier Gutiérrez. Tourists' digital footprint in cities: Comparing big data sources. *Tourism Management*, 66:13–25, 2018.

[41] Gunnar Schröder, Maik Thiele, and Wolfgang Lehner. Setting goals and choosing metrics for recommender system evaluations. In *UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA*, volume 23, page 53, 2011.

[42] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, 2009.

[43] Gagan Deep Sharma, Asha Thomas, and Justin Paul. Reviving tourism industry post-covid-19: A resilience-based framework. *Tourism management perspectives*, 37:100786, 2021.

[44] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. RecVAE: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, jan 2020.

[45] Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference on - WWW '19*. ACM Press, 2019.

[46] Bella Struminskaya, Peter Lugtig, Vera Toepoel, Barry Schouten, Deirdre Giesen, and Ralph Dolmans. Sharing data collected with smartphone sensors: willingness, participation, and nonparticipation bias. *Public opinion quarterly*, 85(S1):423–462, 2021.

[47] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[48] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[49] Quoc-Tuan Truong, Aghiles Salah, and Hady W Lauw. Bilateral variational autoencoder for collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 292–300, 2021.

[50] UNWTO. International tourist arrivals reach 1.4 billion two years ahead of forecasts. *United Nations World Tourism Organization*, 2019.

[51] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web*, pages 391–400, 2017.

[52] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 391–400, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[53] Heitor Werneck, Nícollas Silva, Matheus Viana, Adriano CM Pereira, Fernando Mourão, and Leonardo Rocha. Points of interest recommendations: methods, evaluation, and future directions. *Information Systems*, 101:101789, 2021.

[54] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[55] Cheng Xu, Weimin Wang, Shuai Liu, Yong Wang, Yuxiang Tang, Tianling Bian, Yanyu Yan, Qi She, and Cheng Yang. 3rd place solution to google landmark recognition competition 2021. *arXiv preprint arXiv:2110.02794*, 2021.

[56] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in

location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):1–23, 2016.

[57] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 363–372, 2013.