Neural Computing and Applications (2023) 35:20339–20349
https://doi.org/10.1007/s00521-023-08822-4

ORIGINAL ARTICLE

All-in-one picture: visual summary of items in a recommender system

Pablo Pérez-Núñez[1] · Jorge Díez[1] · Beatriz Remeseiro[1] · Oscar Luaces[1] · Antonio Bahamonde[1]

Received: 4 November 2022 / Accepted: 28 June 2023 / Published online: 20 July 2023

Abstract
Navigation through large volumes of images is a complex and tedious task that requires tools to facilitate the exploration and discovery of visual information. Photo summaries are one of these tools, which consist of selecting a reduced set of images that best represent the original data source. However, creating photo summaries in the context of recommender systems poses several challenges: How to select the most relevant images for each item? How to encode each image? How to evaluate the quality of the generated summary? In this manuscript, we propose a clustering-based method to create a visual summary in the context of a restaurant recommender system, which includes the photos taken by users who visited the restaurants (items) in a given city. These photos are encoded using a deep neural network that takes into account not only their content but also the relationships between users and restaurants. This encoding will allow us to create a visual summary that captures the essence of user tastes and illustrates the gastronomic offer of the city. We also propose a similarity measure between items based on the users who have visited them and an evaluation method that calculates to what extent the summary obtained represents the original data source. The experimentation carried out includes five datasets and the obtained results demonstrate the adequacy of our proposal for the construction of these summaries.

Keywords Recommender systems · Visual summaries · Deep learning, · Clustering

1 Introduction

Complex realities are difficult to assimilate. This is the case of *large volumes of data* that are handled in fields of application such as, for example, Recommender Systems (RS). Valuable information is hidden not only in volume but also behind an intricate web of relationships. Moreover, these data may have a wide variety of types of information that, in addition to quantitative evaluations, may include opinions expressed with texts and/or photos.

In this paper, we present a method to summarize, in a representative and understandable way, the data depicted in these complex scenarios. Particularly, we will focus on datasets with photos of restaurants (taken by customers) in a city. The idea is to *explain*, with a simple visual *summary*, the gastronomic offer of a city that can have tens of thousands of restaurants with hundreds of thousands of photos taken by users.

A first characteristic of the summaries that we are going to present is that they are based on a type of clustering that we could define as *sociological*, different from those that can be conceived based on content. We are not interested in grouping pizzerias or restaurants with certain regional food. Instead, we will consider that two restaurants are *similar* if the sets of users who visited them are also similar. That is, they do not have to offer the same type of food, but they must be *interchangeable in a recommendation* to be visited.

We have taken *visited* as the basic relationship between users and restaurants. Thus, users who consult our visual summaries will have their own tastes and will perceive as

✉ Pablo Pérez-Núñez
pabloperez@uniovi.es

Jorge Díez
jdiez@uniovi.es

Beatriz Remeseiro
bremeseiro@uniovi.es

Oscar Luaces
oluaces@uniovi.es

Antonio Bahamonde
abahamonde@uniovi.es

[1] Artificial Intelligence Center, Escuela de Marina Civil, Universidad de Oviedo, Campus de Gijón, 33204 Gijón, Asturias, Spain

near those restaurants that were visited by similar groups of customers.

Photos that users share after visiting restaurants are considered thoroughly in this research, not only to understand what the users of the restaurant highlighted but also to choose a handful of photos as a summary. This requires a mechanism to understand what the photos mean in our context. Therefore, it is necessary to learn a semantics of photos compatible with restaurants and their relationships with users. And here a second type of similarity appears, that of the photos. We consider two ways of understanding that the photos are similar: if they were taken in the same (or similar) restaurant, or if they look similar from a visual point of view. To implement this measure, which is key in this research work, we have designed a deep neural network that takes these two aspects into account at the same time.

Summaries should allow us to draw a visual panorama of a large volume of complex data. The objective is to facilitate the navigation of users who seek to assimilate a large amount of information. For this reason, the reduced number of images that we are going to select must include, on the one hand, the most relevant aspects and, on the other, photographs that represent the diversity of the whole.

Five datasets taken from the TripAdvisor platform on restaurants in cities of different sizes will be used to evaluate our proposal. However, it is worth noting that the methods that we are going to present could be adapted to other contexts only with slight or even no modifications. The essential issue is to have datasets such as those used in RS; that is, to have users, items, and user reactions to the items expressed through photographs. In the field of tourism, there is a wide variety of possible uses in addition to restaurants, among others: hotels and points of interest (e.g., monuments, landscapes, etc.).

The most relevant contributions of this article are listed below:

- *Item and image encoding* We propose a novel item encoding that uses the set of users who interacted with it. The goal is to have almost the same vector for two restaurants visited by approximately the same set of users. In the case of photos, we have designed a deep neural network to encode them. The idea here is to take into account not only the content of the photos but also the aforementioned relation between users and items.
- *Visual summary of items* We present an automatic system capable of generating, from a large RS dataset (with images), a photo summary that includes the most relevant information. The procedure is divided into two clustering steps for which we have also defined a similarity function consistent with the aforementioned encodings.
- *Evaluation procedure* We pose a method capable of measuring the degree to which the previously created summary can replace the full dataset. The main idea is to check to what extent the photos selected for the final summary allow us to reconstruct the users' behavior.

The rest of the manuscript is organized as follows. After reviewing some related works in Sect. 2, we present the methodology of our proposal in Sect. 3. The experimental setup for evaluating the resulting visual summary is detailed in Sect. 4. Using this procedure, in Sect. 5 we report an exhaustive set of experiments carried out to check the adequacy of our proposal. Finally, Sect. 6 closes the manuscript with the main conclusions.

2 Related work

This work skillfully combines several concepts, approaches, techniques and components, such as visual summaries, deep learning, clustering, and RS. For this reason, this section is dedicated to reviewing some relevant contributions in the most related areas, such as summary algorithms, their evaluation, their use in the specific field of RS, and case studies on restaurant recommendations.

2.1 Summarization

Summary algorithms try to find a small subset of objects (e.g., sentences, images, videos, sounds, etc.) that covers the information of a large set of those objects. The aim is to cover both the diversity of the original set and the representativeness of what is selected as a summary. They must also eliminate redundancies, as it is essential for the summary to be small [1]. When it comes to images, summaries are useful to facilitate navigation through a (usually large) collection of images.

Most of the summarization work found in the literature was done with text documents, see for instance [2]. In this context, algorithms are usually classified as abstractive (they build sentences that summarize the content of the document) or extractive (they select some representative sentences). In the case of images, the abstractive approach does not make sense (except perhaps in very special cases). Selective methods remain and, as when dealing with texts, they include some clustering approaches that require defining the concept of similarity [3].

Regarding the evaluation of summary algorithms, it is worth mentioning that is a controversial issue. In many cases, subjective measures are used, such as carrying out the evaluation through user satisfaction levels or with a relevance score. It has even been claimed that the lack of consensus somehow slows down progress in this field [4].

When summarizing texts, the problem of evaluation is perhaps more complex. The reason is that the semantics of the sentences used in a summary must be compared with that of the original text, which is extremely difficult. An illustrative example of this phenomenon can be found in [4], where the authors present *SummEval*, a set of resources for summarization and evaluation.

An alternative point of view, which appears especially when it comes to summarizing collections of images, is the reconstructive approach. It even also appears when summarizing documents, as in [5]. The idea is to assume that the effectiveness of a summary is reflected in its ability to reconstruct the original set or each individual image of the set [3]. In this case, the images are described by means of a dictionary of objects that can appear in them. In some way, each image is represented as sentences using a bag of words approach. The reconstruction idea is extended to transformer-based encoder and decoder structures, see for instance [6], which deals with multi-document summarization.

2.2 Summarization in recommender systems

As mentioned above, summarizing is closely related to the concept of similarity. The overall idea is to pick one representative element from each group of similar elements. In the context of RS, users and items play a dual role. Therefore, similarities can be employed for both entities.

The similarity of users (or items) involved in RS has been intensively studied. In [7], the authors explored some similarity measures for users. Their target was to determine the set of users that had the same behavior with respect to a given subset of items. For their part, Amer et al. [8] explored the use of the *Jaccard* similarity to improve the performance of an RS. A thorough discussion about combinations of similarity functions devised to improve the performance of an RS can be found in [9]. Unlike previous works, our point of view in this research is to use the similarity to summarize the data collected in the context of RS. Therefore, the similarity will be used here to cluster the available items.

Other interesting works focused on summarization include [10], which uses *Weibo microblogging* data to summarize events using representative texts and images. For this purpose, the authors introduced a co-clustering algorithm to group text and images based on their relationship with users. Finally, Gil et al. [11] introduced *VisualRS*. Their objective was to present the information of an RS in a visual and navigable way, although they do not intend to make a summary.

2.3 Dealing with restaurants

In this paper, we designed a case study for restaurant recommendations. For this purpose, we used TripAdvisor[1] data on restaurants in five cities around the world. It is the largest social network for restaurants, hotels, and tourist activities in general. The photographic information shared on this platform has been previously studied. For example, Giglio et al. [12] used a collection of photographs to understand the perception of luxury by hotel users. More recently, Díez et al. [13] dealt with user photos taken in restaurants and then shared on TripAdvisor. Focusing on authorship, they estimated the probability that a photo was taken by a user. The objective was to provide, along with each personalized recommendation to a user, the photo that was probably taken by that user. The photo would then act as an explanation for the recommendation and would increase the user's interest in the suggestion.

Also in the context of restaurant recommendation, Chu and Tsai [14] presented a hybrid RS. They used a collaborative filtering approach and represented both users and restaurants by means of visual features. A general purpose CNN was employed to extract features from images with additional ad-hoc features. Their key point is the method used to deal with several photos, since the authors used averaging or maximum aggregations instead of a semantic approach as we introduce in this manuscript.

Finally, it is worth quoting [15], which includes a survey focused on the use of *side information* in RS. It is an interesting paper to obtain a general perspective of the topic.

3 Methodology

As stated in Sect. 1, a key element in our approach is to have a set of photos. Users may eventually provide photos and we will use them as a fundamental source of communication. A central reflection is that we understand that users take photos (and share them on a social network) of places that especially attract their attention. Therefore, photos carry an important message about the behavior of the users.

From a formal point of view, our case study includes a set of users $\mathcal{U}$, a set of restaurants $\mathcal{R}$, and a simple relationship between them: *visited*. We could consider other relationships, such as valuation (implicit or explicit); in that case, we would only have to slightly modify the method described in this manuscript.

We will have a dataset $\mathcal{D}$ that contains a triple (u, r, l) for every interaction between any user $u \in \mathcal{U}$ with any

[1] https://www.tripadvisor.com/.

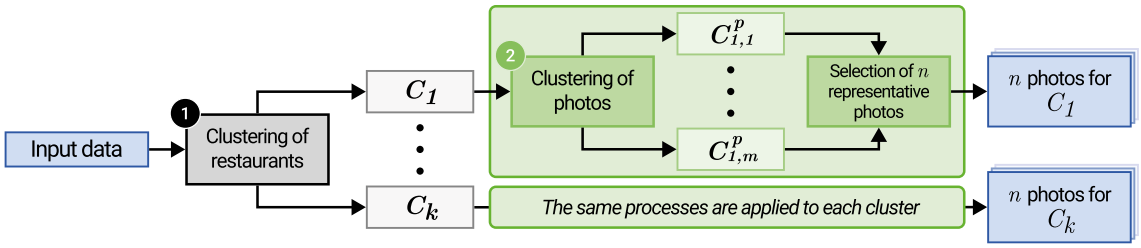

Fig. 1 We propose this workflow to obtain a visual summary: a reduced set of representative photos for each cluster of restaurants, see Sect. 3.1. The photos are encoded using the embedding output, $\hat{\mathbb{Y}}$, provided by the neural network architecture shown in Fig. 2. Finally, the set of photos is chosen from each cluster (wrapped in the green area) following the procedure explained in Sect. 3.2.2. C stands for cluster of restaurants and C^p for cluster of photos, being $C^p_{1,m}$ the m-th cluster of photos in C_1 (Color figure online)

restaurant $r \in \mathcal{R}$, where l is a list of photos $p \in \mathcal{P}$ taken by user u and $\mathcal{P}$ stands for the whole set of photos of the dataset.

All the available information are the data from which an RS is built. Thus, in order to grasp the core idea, it will be especially useful for us to represent each restaurant r by the set of users who visited them. We will call this representation $\mathbb{Y} \in \{0,1\}^{\mathcal{U}}$ and, for a restaurant r, each of its components can be obtained as:

$$\mathbb{Y}^r_i = \text{visited}(u_i, r). \tag{1}$$

Notice that visited(u, r) will return 1 or 0 depending on whether user u visited restaurant r or not.

Figure 1 depicts our proposal to obtain the visual summary, with the following stages: (1) we define a similarity between restaurants to group them and build a hierarchical clustering, and (2) we describe each cluster through a reduced set of photographs that will constitute the intended summary. Each step is detailed in the following sections.

3.1 Clustering of restaurants

To build a cluster, we must first define a *similarity* measure. In this research, we use a function that sets up how interchangeable two restaurants are in a list of recommendations to visit.

More specifically, we use the following definition. For a couple of restaurants, r_1 and r_2, their *similarity* (sim) is given by the dot product of their vectorial representation (1) or, alternatively, the number of users who visited both restaurants. In symbols,

$$\text{sim}(r_1, r_2) = \langle r_1, r_2 \rangle = |r_1 \cap r_2|. \tag{2}$$

Note that this function is different from the *Jaccard similarity*, where the above expression is divided by the cardinal of the union. This is not a good idea in our case, since we propose that similar restaurants (those that are visited by a *similar* set of users) can replace each other in a list of suggestions.

Using this similarity function, the next step involves building a hierarchical clustering [16]. Among all the clusters of similar restaurants obtained, we are only interested in the most outstanding. The clustering is performed with an agglomerative algorithm with *linkage complete*, which stops when the merge of the available groups has a similarity below the 5th% percentile of restaurant similarities. That is, 95% of the similarities yield a cluster merge. Therefore, we try to avoid merging groups of restaurants with little similarity. Note that if all the clusters were joined we would end up with only one group, which would not be informative.

3.2 Clustering of photos

What remains now is to see how to present the restaurants included in the clusters by means of a short list of characteristic photographs of each group of restaurants (see the green area in Fig. 1).

To do this, we need to understand, in a certain sense, the meaning of the photos, and select the most representative of each group of similar restaurants. These two steps are following described in depth.

3.2.1 Photography embedding

We will use an embedding of the photos in a Euclidean space to assign meaning to the images. The idea is to learn an embedding that represents each photo as a point close to those assigned to other photos of the same restaurant and to similar photos from a visual point of view.

To implement this objective, we devised a deep neural network that aims at detecting the restaurant where each photo was taken. The network learns an embedding

$$\text{Embedding} : \mathcal{P} \longrightarrow \hat{\mathbb{Y}} \cong \mathbb{Y}, \tag{3}$$

where $\mathcal{P}$ represents the set of photos and $\hat{\mathbb{Y}}$ is the prediction of the neural network. Let us recall that restaurants are

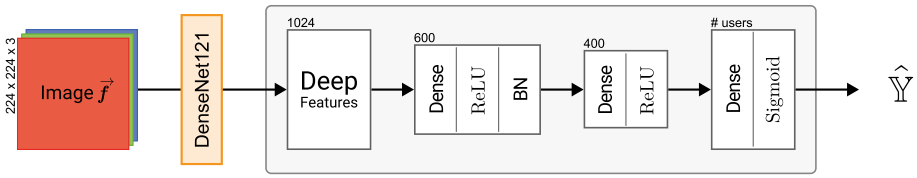

Fig. 2 Network used to learn the embedding (3) from photographs to the set of estimations of codes of restaurants, $\hat{\mathbb{Y}}$

represented by a binary vector that encodes the set of users who visited them, (1). Thus, this embedding will project each photo to a vector space where the *i*-th component of each vector is the probability that such photo had been taken in a restaurant visited by the corresponding user, u_i. In the rest of the manuscript, we will use $\hat{\mathbb{Y}}$ to refer to the entire method presented.

At this point, we can see that the representation of restaurants and the way we define their similarity (2) is very important. The embedding assignment may be not accurate enough to precisely predict the restaurant where a photo was taken. However, we expect that the embedding will associate the photo to a point at $\{0,1\}^{\mathcal{U}}$ very close, not only to its restaurant but also to other similar restaurants; that is, to those of its cluster (see Sect. 3.1).

Returning to the definition of the embedding (3), from a formal point of view, it can be seen as a *multi-label* classifier. Fig. 2 depicts the deep neural network used to build this function. This network first applies the convolutional base of a DenseNet [17], pre-trained on the ImageNet dataset [17], to convert an input RGB image into a 1024-feature vector. The rest of the architecture is composed of fully connected (Dense) layers of different sizes, along with rectified linear unit (ReLU) [18], and batch normalization (BN) [19] layers. Finally, a sigmoid activation function is applied to obtain the estimation of restaurant codification.

3.2.2 Selection of photographs

Let C be a cluster of restaurants obtained following the procedure introduced in Sect. 3.1 and $\mathcal{P}^C \subset \mathcal{P}$ the set of photographs of the restaurants in C. Among all the photos in $\mathcal{P}^C$, we select the most representative ones employing the same hierarchical method used to do the restaurant clustering, with the same parameters: an agglomerative hierarchical clustering with *linkage complete*. As before, the algorithm will stop when the merging of available clusters has a similarity less than the 5th percentile of the similarities in $\mathcal{P}^C$.

Among all the m photo clusters obtained in this stage, we will consider the n clusters with more elements (see Fig. 1). Next, in each cluster, we will select the most similar photo (using the dot product, as in (2)) to the centroid. This procedure will result in a set of n representative photos of C, which we will call $\mathcal{S}^C$. Finally, we will repeat this process for the k restaurant clusters, thus obtaining the visual summary, $\mathcal{S} \subset \mathcal{P}$, for the entire dataset.

4 Experimental setup

This section describes the experiments carried out to evaluate the performance of our approach. We first present the five datasets collected from TripAdvisor, which include the users' reviews from different cities. Then, we detail the evaluation procedure and the experimental process.

4.1 Datasets

We downloaded users' reviews of restaurants located in five cities around the world[2] collected by TripAdvisor during 2018 and 2019. Each review contains between zero (no photos attached by the user) and four images (maximum shown by the platform).

The selection of cities was made with the purpose of including a range of sizes, in terms of restaurants (which are, obviously, highly correlated to the population). In particular, we used three Spanish cities: Barcelona (population: 1.6 million) and Madrid (pop.: 3.2 million), the two largest in the country; and Gijón, a medium size city of around 300,000 inhabitants. We also used data from other big cities of the world, such as New York City (8.3 million), and Paris (2.1 million). Table 1 shows the figures of each dataset after applying some pre-processing explained below.

As can be seen in Fig. 3, the raw data was cleaned up by applying some filters. First, we kept only the most recent review for each pair user/restaurant. That is, if a user reviewed a restaurant several times, we filtered out the oldest reviews. We also removed all photos without food. To build the vectorial representation (1) of the restaurants, instead of using the complete set of users $\mathcal{U}$ (it would be unfeasible due to its size), we used a sample with the 25% more active users (those with the largest number of reviews). Next, we eliminated restaurants with less than five users (from those used for encoding) and also those that did not have at least five photos. The clustering of

[2] The datasets are available for download at Zenodo [20].

Table 1 Basic statistics of the TripAdvisor datasets used in our experiments

	All data				Training data			
	#Users	#Rest	#Reviews	#Photos	#Users	#Rest	#Reviews	#Photos
Gijón	4203	373	6475	14,241	2297	373	3917	8748
Barcelona	25,230	3236	49,024	105,638	14,339	3236	32,229	70,842
Madrid	33,222	3707	65,191	141,649	18,512	3707	41,714	92,287
New York city	43,581	3733	74,806	130,992	24,147	3733	46,392	82,886
Paris	46,794	6764	88,406	171,296	27,648	6764	60,260	118,017

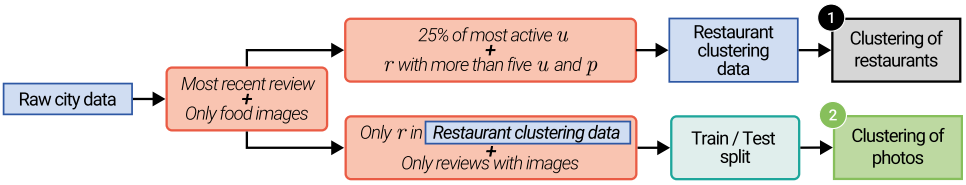

Fig. 3 Dataset filtering and partitioning procedure for each of the two stages shown in Fig. 1. The filters applied are inside the red boxes (Color figure online)

restaurants described in Sect. 3.1 was carried out with the resulting data.

Then, to perform the clustering of photos, we have to start again from the dataset with the first two filters applied (most recent review + only food images). From this one, we keep only the restaurants that appear in the set created for the previous phase and eliminate the reviews without images. By doing this separate procedure for the second stage, we will have the images of all the users available and not only the ones from the 25% more active users. As explained before, we will estimate the ability of our approach to summarize the gastronomical offer of a city. Thus, we split the data with respect to the users, so that the information in the test set corresponds to users never seen during the training stages. By doing so, we will be emulating cold-start situations, which are very common in the field of recommender systems. The split was made to retain approximately 50% of the users in each partition, ensuring that all restaurants that appear in the triplets of the test set also appear in the training set.

4.2 Evaluation

This section describes how to assess the quality of the visual summary $\mathcal{S}$. To do that, we will measure the degree to which this summary can replace the entire collection of photos $\mathcal{P}$. The central idea is to measure to what extent the visual summary allows us to reconstruct the behavior of users that, in this context, will be the list of restaurants they visited.

To give a precise formulation, we will consider that each user u is described as a set of photographs $\mathcal{P}^u \subset \mathcal{P}$. The reconstruction will be done in two steps. The first is to determine the photo in the summary with the maximum similarity to one of the photos taken by the user u:

$$j* = \underset{i,j}{\operatorname{argmax}}\ \operatorname{sim}(\mathcal{P}^u_i, \mathcal{S}_j), \quad i = 0..|\mathcal{P}^u|, j = 0..|\mathcal{S}| \tag{4}$$

where sim is the function defined in (2).

The second step consists in associating the user to the cluster of restaurants where the photo $\mathcal{S}_{j*}$ was taken, which we will call C^u. According to the summary, we will understand that the user's habits include the restaurants in that cluster. Let us recall that clusters are built in such a way that their components can be interchangeable in a list of recommendations to visit (1), (2). Therefore, it seems reasonable to accept a cluster as a useful description of users' behavior.

Finally, we define the quality of the summary as the proportion of users for whom C^u contains at least one of the restaurants visited by u. In symbols,

$$\mathcal{R}^{C^u} \cap \mathcal{R}^u \neq \emptyset. \tag{5}$$

4.3 Experiment description

All methods considered in the experimentation start from a set of restaurant clusters obtained as explained above.

Table 2 Different approaches for the assignment of a cluster of restaurants to a user. The baseline approach assigns the cluster with the largest number of photos, so it does not depend on their encoding

	Photo encoding	Cluster selection
$\hat{\mathbb{Y}}$ (Our proposal)	Network output	Highest inner product
$\hat{\mathbb{Y}}_{rnd}$	Network output	Random selection
$\mathbb{D}$	DenseNet vector	Closest (Euclidean distance)
$\mathbb{D}_{rnd}$	DenseNet vector	Random selection
$\mathbb{B}$ (Baseline)	N/A	Largest cluster

Therefore, the evaluation is focused on the final visual summary.

Our proposal ($\hat{\mathbb{Y}}$) summarizes each group of restaurants by selecting n images. Then, each user in the test set is assigned the group with the most similar photo to those provided by the user, (4). We compared the performance of $\hat{\mathbb{Y}}$ with several variants obtained by ablating its two main components; that is, removing the network to encode the photos and using a CNN state-of-the-art encoding approach ($\mathbb{D}$), replacing the cluster selection method with a random choice ($\hat{\mathbb{Y}}_{rnd}$), or applying both modifications ($\mathbb{D}_{rnd}$). Table 2 summarizes the components of all these approaches.

On the other hand, we also tested a baseline approach ($\mathbb{B}$) that simply selects the cluster of restaurants with the largest number of photos in the training set, thus not depending on any image encoding. In a sense, this method uses a kind of *popularity* measure to assign the cluster of restaurants.

In order to test the robustness of our approach we tried a range of values for some parameters of the experiments. Thus, we run the experiments considering a list of n photos to represent each cluster, where $n \in [1..5]$.

With respect to the training of $\hat{\mathbb{Y}}$, we used a grid search on the training dataset of Barcelona that yielded a learning rate $\alpha = 5 \cdot 10^{-4}$ with linear decay down to $1 \cdot 10^{-5}$, a batch size $b = 1024$, and the weights for the weighted loss $w_0 = 1$ and $w_1 = 5$. The network was trained using an early stopping strategy with a maximum of 4000 epochs.

5 Results: analysis and discussion

This section presents the results obtained during the experimentation carried out on the five TripAdvisor datasets. More specifically, Table 3 reports a detailed comparison of the different methods evaluated on the five cities studied. Remember that the evaluation method considers a successful case (hit) when the assigned cluster contains at least one restaurant visited by the user, (4). To ease the reading of these results, we have expressed them as percentages.

The first column references the name of the city. To check the robustness of the procedure, we have distinguished the scores obtained with users who have at least one photo in the test set (≥ 1), at least two (≥ 2), three (≥ 3), or four (≥ 4). This is indicated in the second column ($\#t$) of the table. The results for the baseline method ($\mathbb{B}$) are displayed in the third column. The rest of the table is split into two parts: (1) the left-hand side shows the scores obtained when the photos were selected using the clustering method, and (2) the right part reproduces the results when the photos were randomly chosen. As stated in Sect. 4.3, we have also varied the number n of representative photos for each cluster, ranging from 1 to 5 (numbered columns) to check how it affects the performance.

The first thing that stands out is the improvement of every model as the number of representative images (n) per cluster of restaurants increases. As expected, the greater the number of representative images, the greater the probability of correctly assigning a test user to her most suitable cluster of restaurants, (4). The exception is the $\mathbb{B}$ model that chooses the largest cluster without further consideration and is, therefore, independent of the number of photos in the summary. This expected improvement in results can also be observed in the other direction (rows), when the number of images per user ($\#t$) increases within the same city.

Comparing the representative photo selection strategies (the two main parts of the table), the difference in results is quite noticeable. The random strategy appears to be the worst option when selecting images to represent a cluster of restaurants. Thus, the need for a strategy such as the one proposed in this paper is more than justified. The results obtained using photo clustering are always better than the corresponding ones on the right-hand side of the table. In some cases, like Madrid ≥ 4 with $n = 5$ using $\hat{\mathbb{Y}}$, the difference is remarkable (26%). It is worth mentioning the case of New York City, given that is the one with less difference between both strategies, particularly when n is five.

Focusing now on the model comparison, it is observed that, regardless of the photo selection strategy, there are two main behaviors that stand out. The first one is the surprisingly good results of the majority model ($\mathbb{B}$), driven by the fact that in some cities the visits of the customers are not uniform. This is the case of Gijón (due to its small size) and Madrid (with a lot of tourists who follow the advice of the guides and visit the same places). The good performance of $\mathbb{B}$ in those cities can clearly be seen in Fig. 4 (in blue), where all the results in Table 3 are graphically represented using radial charts.

The second noteworthy behavior is that our model ($\hat{\mathbb{Y}}$) outperforms the DenseNet encoding alternative ($\mathbb{D}$) in all the performed tests. There is only one case where this does not happen, and that is Madrid ≥ 4 with $n = 1$ for the random selection strategy. In this case, the $\mathbb{D}_{rnd}$ beats $\hat{\mathbb{Y}}_{rnd}$ by only 0.2%, which does not seem relevant. On the opposite side is the case of Gijón ≥ 4 with $n = 3$, in which our model improves the DenseNet alternative by a remarkable 60%.

In order to verify if this difference in favor of our model is statistically significant, we carried out a Bonferroni-Dunn test with $\alpha = 0.05$. The results of the test,

Table 3 Experimental results in percentage (see Sect. 4.2). The scores were obtained with the data of the five cities indicated in the leftmost column. To check the robustness of the procedure, we split the test elements by the number of photos available of each user (column $\#t$), as well as by the number of representative photos for each cluster, varying from 1 to 5 (numbered columns). We also compared two methods for the selection of the representative images in each cluster of restaurants: by clustering of images (left half of the table) and by random selection (right half). The results obtained in each configuration for these approaches are labeled as: the baseline procedure ($\mathbb{B}$), using clustering or random selection of images encoded with DenseNet ($\mathbb{D}$ and $\mathbb{D}_{rnd}$, respectively), or encoded by our proposed neural network ($\hat{\mathbb{Y}}$ and $\hat{\mathbb{Y}}_{rnd}$ respectively); see Sect. 4.3

			Using photo clustering										Random selection of photos									
			1		2		3		4		5		1		2		3		4		5	
	$\#t$	$\mathbb{B}$	$\mathbb{D}$	$\hat{\mathbb{Y}}$	$\mathbb{D}$	$\hat{\mathbb{Y}}$	$\mathbb{D}$	$\hat{\mathbb{Y}}$	$\mathbb{D}$	$\hat{\mathbb{Y}}$	$\mathbb{D}$	$\hat{\mathbb{Y}}$	$\mathbb{D}_{rnd}$	$\hat{\mathbb{Y}}_{rnd}$	$\mathbb{D}_{rnd}$	$\hat{\mathbb{Y}}_{rnd}$	$\mathbb{D}_{rnd}$	$\hat{\mathbb{Y}}_{rnd}$	$\mathbb{D}_{rnd}$	$\hat{\mathbb{Y}}_{rnd}$	$\mathbb{D}_{rnd}$	$\hat{\mathbb{Y}}_{rnd}$
Gijón	≥ 1	36.1	3.4	35.3	6.3	33.6	9.5	39.6	11.2	41.3	12.6	42.0	2.8	20.1	2.5	27.4	3.1	29.7	3.1	31.1	3.7	32.0
	≥ 2	58.6	6.6	53.2	10.0	50.5	11.2	57.7	13.9	59.2	15.7	59.2	2.1	34.7	2.4	44.1	3.9	45.6	4.2	46.8	4.2	48.0
	≥ 3	69.6	9.6	65.6	12.0	64.0	12.8	70.4	18.4	71.2	20.0	71.2	3.2	48.0	4.0	55.2	5.6	56.8	4.0	56.8	4.8	56.8
	≥ 4	73.8	15.4	72.3	16.9	70.8	16.9	76.9	18.5	76.9	21.5	76.9	3.1	55.4	3.1	58.5	4.6	56.9	3.1	56.9	4.6	56.9
Madrid	≥ 1	4.6	1.9	9.7	2.8	12.9	3.3	14.2	3.8	15.1	4.1	15.3	1.0	3.3	1.2	6.8	1.5	8.3	1.8	9.6	1.9	10.4
	≥ 2	10.5	2.6	16.1	3.8	22.1	4.2	24.2	4.6	25.8	5.1	26.1	1.4	5.8	1.6	12.7	2.1	14.4	2.4	17.0	2.7	18.3
	≥ 3	14.8	2.9	19.1	3.9	27.2	4.1	29.7	4.5	31.9	5.5	31.9	1.8	7.2	1.8	16.0	2.3	16.9	2.7	19.9	3.3	21.4
	≥ 4	17.4	3.8	20.6	4.7	30.4	5.2	34.3	5.5	37.0	6.5	37.2	0.9	7.3	1.4	16.8	2.4	17.9	2.8	22.2	3.8	23.1
NYC	≥ 1	12.0	2.3	9.4	3.5	13.2	4.1	14.5	4.6	15.2	4.9	16.1	0.8	1.2	0.9	6.3	1.1	5.5	1.6	5.3	1.8	5.5
	≥ 2	22.5	3.7	16.3	5.0	22.6	5.8	24.2	6.3	25.4	6.6	27.2	0.9	1.2	0.8	10.6	0.9	9.5	1.8	9.4	2.2	9.4
	≥ 3	31.1	4.8	19.3	6.6	26.8	7.0	29.9	7.8	31.5	8.1	33.9	1.0	1.3	0.9	12.2	1.2	11.4	1.9	11.4	2.4	11.2
	≥ 4	36.0	6.6	21.8	8.4	30.8	8.9	33.9	9.2	36.0	9.6	38.7	1.3	1.1	1.2	14.0	1.3	13.0	1.8	13.0	2.1	12.7
Barcelona	≥ 1	9.6	3.8	13.0	4.5	16.4	5.4	18.1	6.2	18.0	6.7	18.6	1.6	8.6	1.8	10.7	2.0	13.1	2.1	12.8	2.2	16.2
	≥ 2	17.6	5.2	20.3	5.8	24.9	6.7	26.5	7.1	25.8	7.7	26.8	1.9	14.6	1.8	17.8	2.1	21.5	2.4	20.9	2.4	26.3
	≥ 3	22.9	5.8	25.3	6.4	30.9	7.7	32.9	8.2	31.8	9.0	32.7	2.0	17.7	2.0	21.3	2.5	25.1	2.7	24.8	3.0	31.0
	≥ 4	28.0	5.7	29.4	6.1	34.9	7.5	36.6	8.0	35.3	8.7	35.9	2.0	21.5	2.5	25.6	3.2	29.2	3.3	29.0	3.8	35.5
Paris	≥ 1	3.0	1.8	9.4	2.5	10.9	2.9	11.8	3.1	11.9	3.3	12.2	0.7	2.6	0.9	5.8	1.1	6.7	1.4	7.5	1.7	7.7
	≥ 2	6.7	2.7	17.0	3.5	19.0	3.7	20.9	3.9	21.1	3.9	21.6	0.7	4.0	1.0	10.8	1.0	11.3	1.2	12.7	1.6	12.9
	≥ 3	9.4	3.2	21.6	3.8	24.3	3.9	26.7	4.2	26.9	4.3	27.7	0.8	4.4	0.9	13.7	0.9	14.0	1.0	15.9	1.4	15.6
	≥ 4	12.7	3.2	25.9	3.8	29.1	3.8	32.2	4.0	32.1	4.1	33.8	0.5	4.3	0.8	16.5	0.9	17.4	0.9	18.7	1.8	18.3

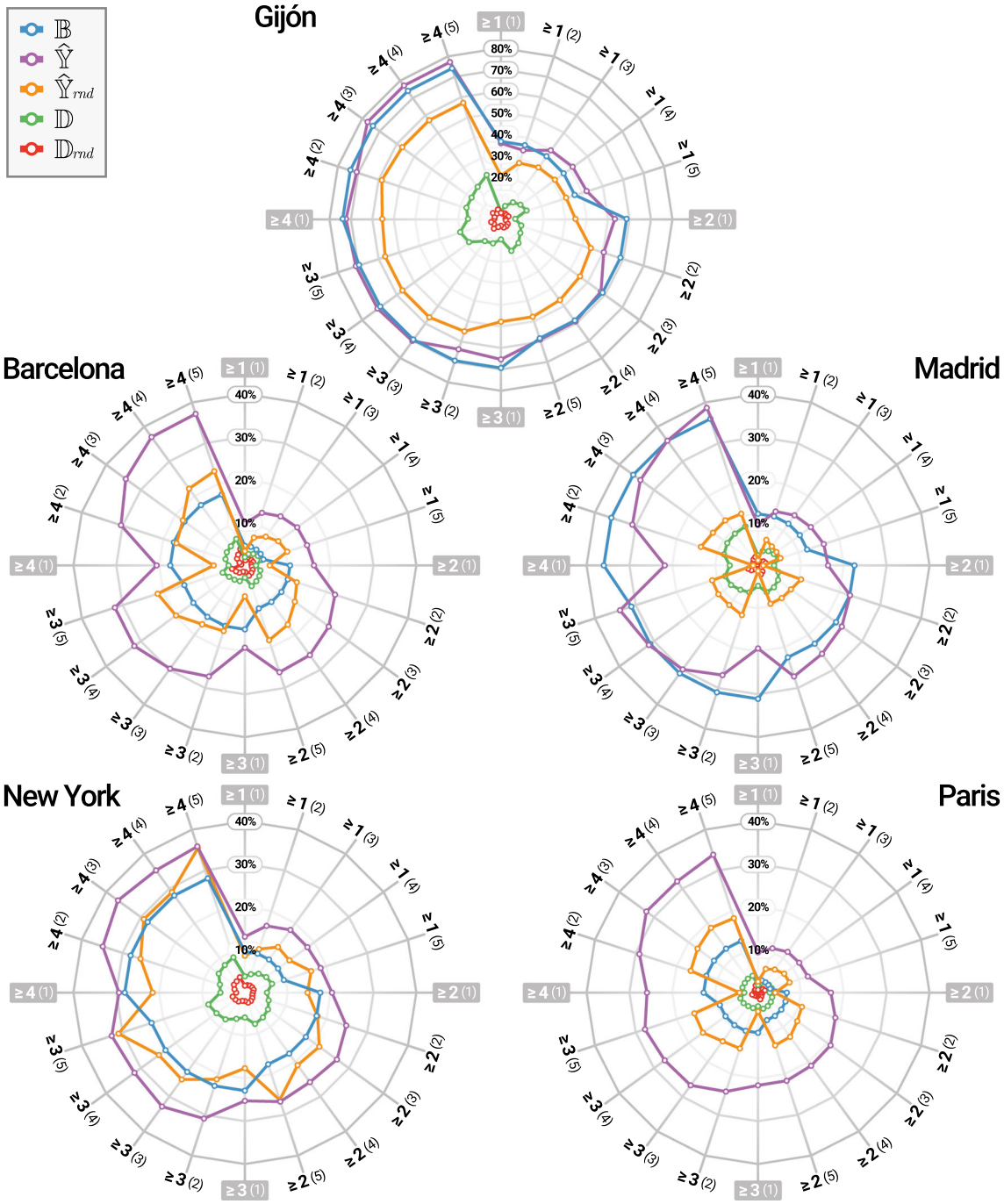

Fig. 4 Radar charts of the results of each city. The axes represent the number of photos. For example, "$\geq$ 2 (3)" stands for the scores achieved for users with at least 2 photos and where summaries were built with 3 photos of each cluster. Notice that Gijón has a maximum of 80%, while the rest of the cities have 40%

graphically depicted in Fig. 5, indicate that $\hat{\mathbb{Y}}$, our model, is significantly better than the other four models by a wide margin. Furthermore, its random version ($\hat{\mathbb{Y}}_{rnd}$) outperforms traditional encoding in both of its two configurations ($\mathbb{D}$ and $\mathbb{D}_{rnd}$). Regarding the baseline ($\mathbb{B}$), despite being the second best model slightly above $\hat{\mathbb{Y}}_{rnd}$, there is no significant evidence to say that it is statistically superior.

The scores show that the image encoding is of crucial importance for the task at hand. Our proposed encoding, $\hat{\mathbb{Y}}$, maps each photo into a space taking into account the users who visited the restaurant where it was taken. In a sense, the model obtained by our deep neural network generalizes the latent features of the gastronomic offer of the restaurant, and that made a specific group of users to visit it.

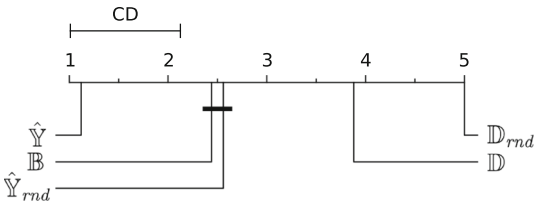

Fig. 5 Bonferroni–Dunn test with $\alpha = 0.05$

The DenseNet encoding ($\mathbb{D}$), on the contrary, has nothing to do with the taste of users. The reason is that it is an encoding devised to achieve good performance in general purpose computer vision tasks, such as object recognition. We used the DenseNet as a starting point but our posterior processing has proven to be essential in order to achieve an adequate summary of representative images of restaurants regarding the users' tastes.

6 Conclusions

This paper presents a method to visually summarize the information of an RS dataset. In this case, the interactions between users and items include photos that play a key role. In fact, the summaries built are a reduced set of photos that contain the condensed information from the dataset. To illustrate the proposed method, we used a case study with five restaurant datasets taken from TripAdvisor. In this context, the visual summary is a short description, in a few photos, of the gastronomic offer of a city.

The key piece of the proposal is the encoding of users' photos. For this purpose, we use a deep neural network that takes into account not only the visual characteristics of the photos but also the relationship between the restaurants where they were taken and the users who visited them.

When dealing with summaries, it is not trivial to establish the evaluation method that should be used to measure their quality. In this research, we chose to contrast the ability of the summaries to be able to reconstruct and generalize the gastronomic behavior of the users. Regarding the experimentation carried out, we designed an ablation study to analyze the relevance of the different components of the proposed method. The result is that performance plummets if we skip any of the steps detailed in the manuscript.

The approach introduced can be useful in the treatment of RS datasets with multimedia elements that arise from the interaction between users and items. Additionally, the definition of similarity used in this research, which is the centerpiece, can be extended to other types of data with relative ease.

While this research presents an innovative approach to creating visual summaries, there are some limitations and opportunities for future work. For instance, the evaluation of visual summary quality is solely based on the ability to reconstruct and generalize the users' behavior, which may not be sufficient in some circumstances. Additionally, the computational cost of this approach could be a downside for some for practical applications.

Future opportunities include exploring the performance of our approach in other datasets with a similar structure (e.g., Amazon reviews). It could be also interesting to extend the approach to handle other types of multimedia elements, such as videos or audio files, in order to build more complete summaries. Different similarity metrics and clustering algorithms could be also investigated to improve the quality of visual summaries.

Acknowledgements This work was funded under grant PID2019-109238GB-C21 from the Spanish Ministry of Science and Innovation, partially supported with ERDF funds. Pablo Pérez-Núñez acknowledges the support of the Principado de Asturias Regional Government under *Severo Ochoa* predoctoral program (ref. BP19-012). We are grateful to NVIDIA Corporation for the donation of the Titan Xp GPUs used in this research.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availibility The datasets used in this research are available in the Zenodo repository: https://doi.org/10.5281/zenodo.5644892.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest directly or indirectly related to the presented work.

References

1. Nenkova A, McKeown K et al (2011) Automatic summarization. Foundations and Trends® in Information Retrieval 5(2–3):103–233
2. Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. Artif Intell Rev 47(1):1–66
3. Yang C, Shen J, Peng J, Fan J (2013) Image collection summarization via dictionary learning for sparse representation. Pattern Recognit 46(3):948–961
4. Lloret E, Plaza L, Aker A (2018) The challenging task of summary evaluation: an overview. Lang Resour Eval 52(1):101–148

5. Chu E, Liu P (2019) Meansum: a neural model for unsupervised multi-document abstractive summarization. In: International conference on machine learning, pp 1223–1232
6. Li W, Xiao X, Liu J, Wu H, Wang H, Du J (2020) Leveraging Graph to Improve Abstractive Multi-Document Summarization. In: 58th Annual Meeting of the Association for Computational Linguistics, pp. 6232–6243
7. Gazdar A, Hidri L (2020) A new similarity measure for collaborative filtering based recommender systems. Knowled-Based Syst 188:105058
8. Bag S, Kumar SK, Tiwari MK (2019) An efficient recommendation generation using relevant Jaccard similarity. Inf Sci 483:53–64
9. Amer AA, Abdalla HI, Nguyen L (2021) Enhancing recommendation systems performance using highly-effective similarity measures. Knowled-Based Syst 217:106842
10. Qian X, Li M, Ren Y, Jiang S (2019) Social media based event summarization by user-text-image co-clustering. Knowled-Based Syst 164:107–121
11. Gil S, Bobadilla J, Ortega F, Zhu B (2018) VisualRS: Java framework for visualization of recommender systems information. Knowled-Based Syst 155:66–70
12. Giglio S, Pantano E, Bilotta E, Melewar T (2020) Branding luxury hotels: evidence from the analysis of consumers' big visual data on TripAdvisor. J Bus Res 119:495–501
13. Díez J, Pérez-Núñez P, Luaces O, Remeseiro B, Bahamonde A (2020) Towards explainable personalized recommendations by learning from users' photos. Inf Sci 520:416–430
14. Chu W-T, Tsai Y-L (2017) A hybrid recommendation system considering visual information for predicting favorite restaurants. World Wide Web 20(6):1313–1331
15. Sun Z, Guo Q, Yang J, Fang H, Guo G, Zhang J, Burke R (2019) Research commentary on recommendations with side information: a survey and research directions. Electron Commer Res Appl 37:100879
16. Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244
17. Huang G, Liu Z, Weinberger KQ, van der Maaten L (2017) Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition, pp 4700–4708
18. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: 27th International conference on machine learning, pp 807–814
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456
20. Pérez-Núñez P, Luaces O, Díez J, Remeseiro B, Bahamonde A (2021). TripAdvisor Restaurant Reviews Zenodo. https://doi.org/10.5281/zenodo.5644892