Learning User Representations for Open Vocabulary Image Hashtag Prediction – Supplementary

Thibaut Durand
Borealis AI Simon Fraser University
thibaut.durand@borealisai.com

A. Supplementary

A.1. Dataset

In this section, we give more information about the datasets. The dataset statistics are shown in Table 1. We limit the maximum number of images per user to 200 to prevent that the dataset biased to a small number of dominant users. We also limit minimum number of images per user to 50 to have enough information to extract a user representation. For the fixed vocabulary dataset, we define the vocabulary as the set of hashtags that are used at least 50 times by at least 3 unique users. Note that the fixed vocabulary dataset has less users and images because a lot of images are ignored because they do not have at least one valid hashtag. The open vocabulary dataset is more challenging than the fixed vocabulary dataset because there are more hashtags and the dataset is highly imbalanced. The Figure 1 shows the number of images per hashtag, the number of unique users per hashtag and the number of images per user for each dataset. The Figure 2 shows the word cloud representation of the hashtag distribution on the training set of the open vocabulary dataset.

	TRAIN	VAL	TEST
OPEN VOCABULARY			
num users	21,441	3,070	6,130
avg images per user	119	119	119
avg hashtags per image	4.49	4.46	4.49
num hashtags	442,054	487,454	568,883
FIXED VOCABULARY			
num users	14,574	2,042	4,066
avg images per user	111	113	110
avg hashtags per image	3.85	3.69	3.67
num hashtags	18,583	-	-

Table 1. Dataset statistics.

A.2. Metrics

The models are evaluated with three different metrics: Accuracy@k, Precision@k and Recall@k. We note $Rank(x, u, k)$ the set of top k ranked hashtags by the model for image x and user u, and $GT(x, u)$ the set of hashtags tagged by the user u for the image x.

- **Accuracy@k (A@k)**. The Accuracy@k measures how often at least one of the ground-truth hashtags appears in the k highest-ranked predictions.

$$A@k = \sum_{i=1}^{N} \frac{\mathbb{1}[Rank(x_i, u_i, k) \cap GT(x_i, u_i) \neq \emptyset]}{N} \tag{1}$$

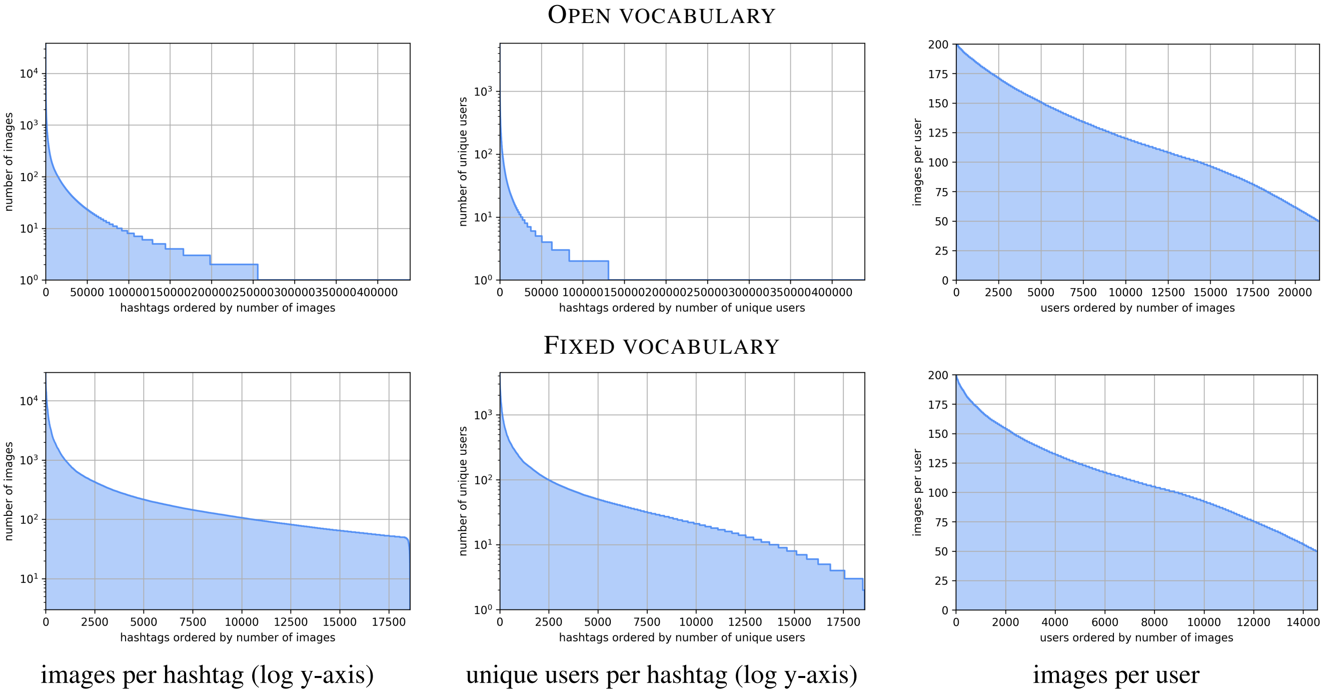


Figure 1. Dataset analysis. For each dataset, we show the number of images per hashtag, the number of unique users per hashtag and the number of images per user. We observe that the open vocabulary dataset is highly imbalanced.

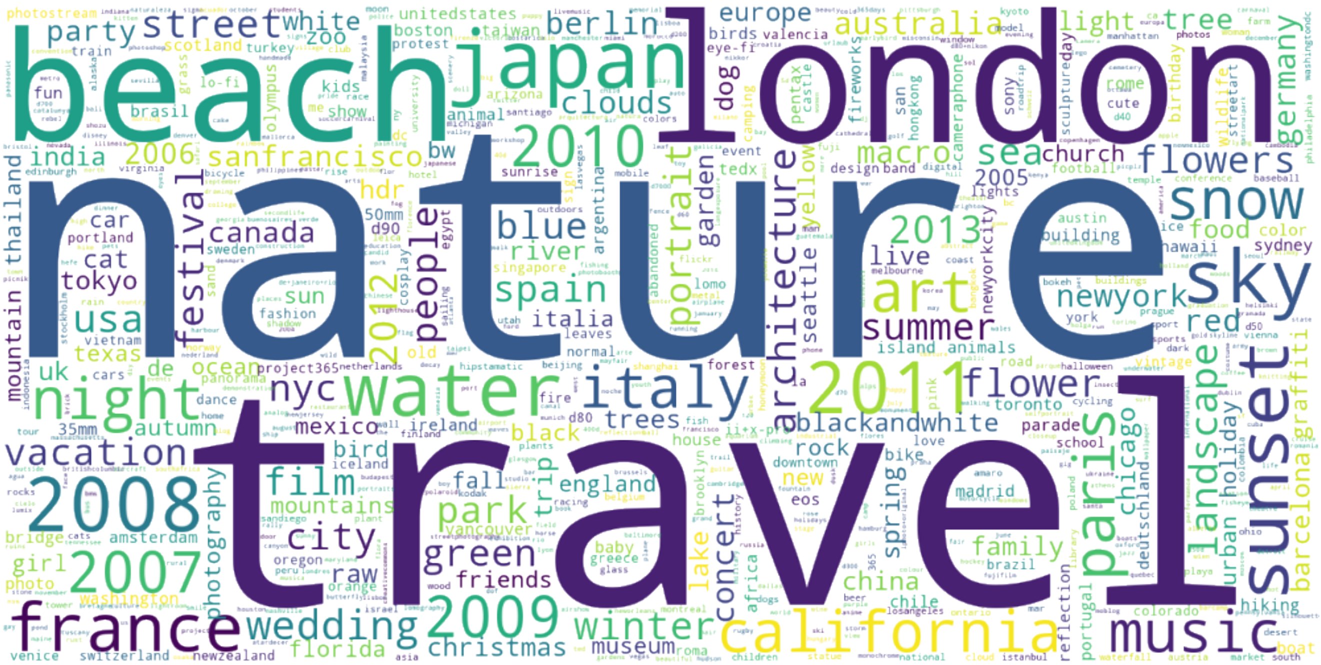


Figure 2. Word cloud representation of the hashtag distribution on the open vocabulary dataset (training set).

- **Precision@k (P@k)**. The Precision@k computes the proportion of relevant hashtags in the top-k predicted hashtags. $HR(x_i, u_i)$ is the rank of the positive hashtag with the lowest score. We use this definition because a lot of images have less than 10 hashtags.

$$P@k = \frac{1}{N} \sum_{i=1}^{N} \frac{|Rank(x_i, u_i, k) \cap GT(x_i, u_i)|}{\min(k, HR(x_i, u_i))} \tag{2}$$

- **Recall@k (R@k)**. The Recall@k computes the proportion of relevant hashtags found in the top-k predicted hashtags.

$$R@k = \frac{1}{N} \sum_{i=1}^{N} \frac{|Rank(x_i, u_i, k) \cap GT(x_i, u_i)|}{|GT(x_i, u_i)|} \tag{3}$$

The Accuracy@1 and Precision@1 are equivalent by defintion.

A.3. Model architecture for the fixed vocabulary setting

For the fixed vocabulary dataset, we made some changes to the model proposed in section 3 of the paper to have a model more similar to [1]. The model is shown in Figure 3. The main difference is that we do not use pretrained word embeddings. Instead, we learn an embedding per hashtag as in [1]. We use a fully-connected layer to predict the hashtags from the user conditional visual representation.

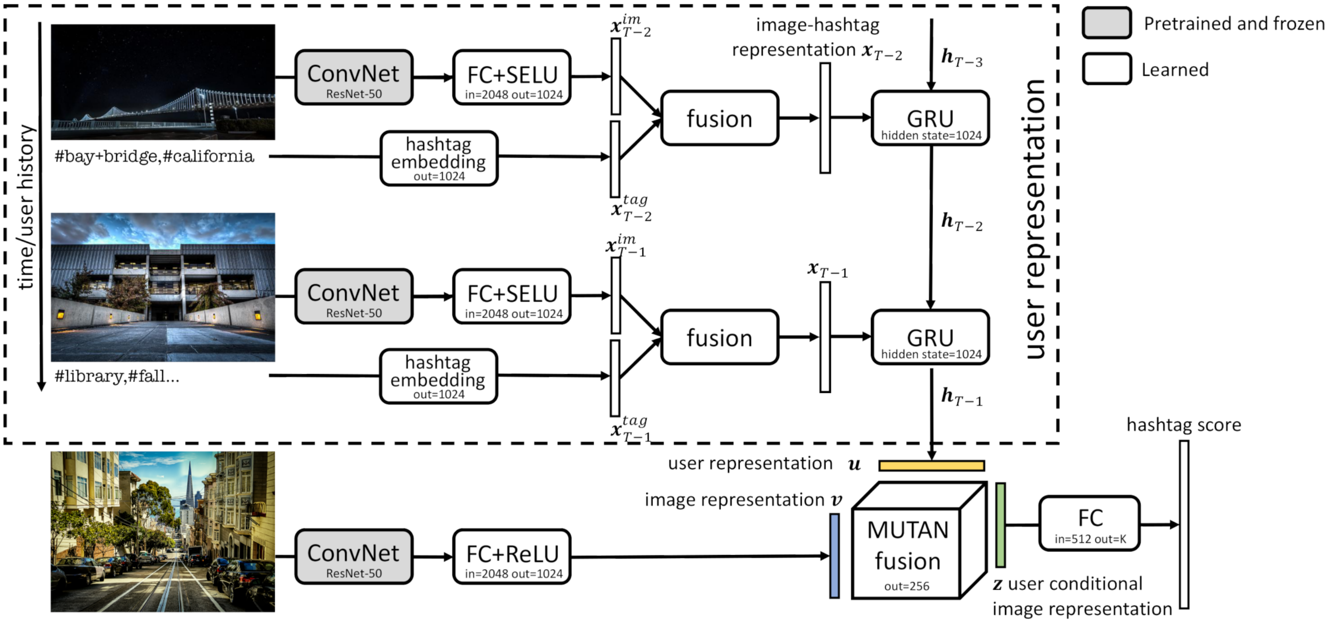
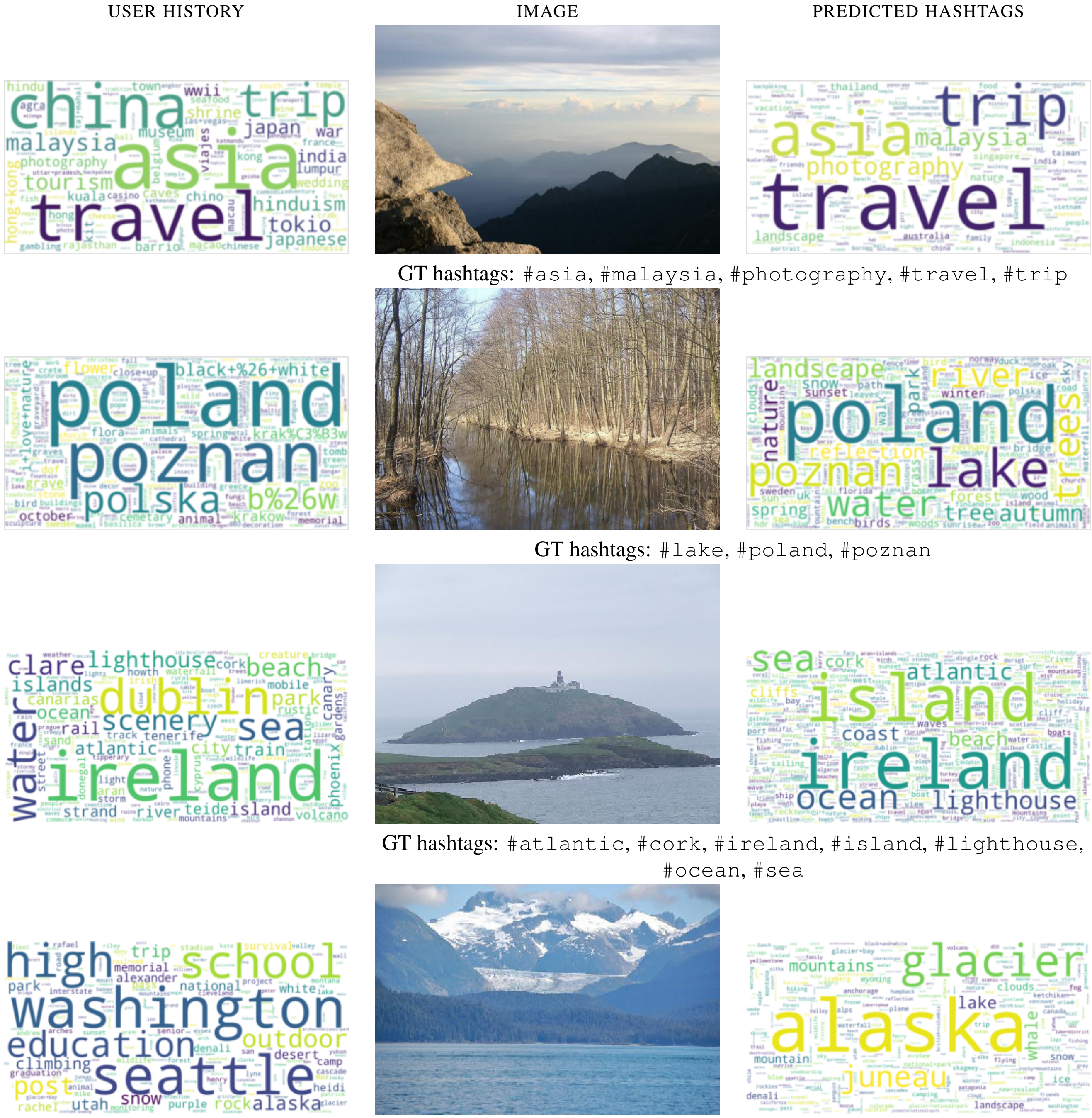


Figure 3. The variant of our hashtag prediction model for the fixed vocabulary setting. The main difference is that model does not use pretrained word embeddings.

A.4. Image-to-hashtag retrieval

In this section, we give some qualitative results of our model for the image-to-hashtag retrieval. In Figure 4, we show the predicted hashtags given the image and the user representation. The left image shows the word cloud representation of the user history. We can see that it is very difficult to predict the hashtags (particularly geographical hashtags) without the user history.


USER HISTORY

IMAGE

PREDICTED HASHTAGS

GT hashtags: `#asia`, `#malaysia`, `#photography`, `#travel`, `#trip`

GT hashtags: `#lake`, `#poland`, `#poznan`

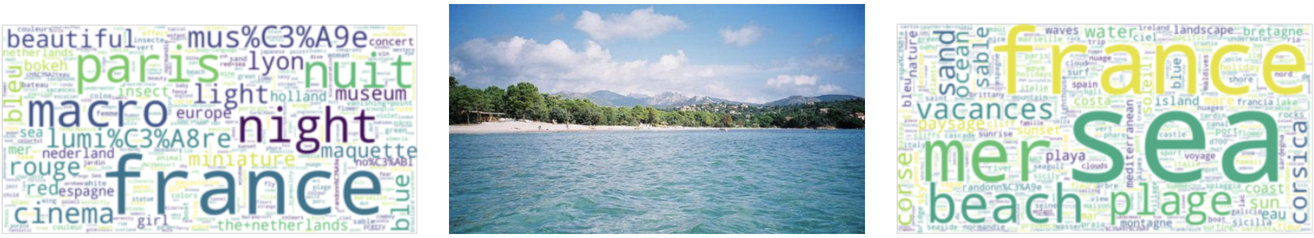
GT hashtags: `#atlantic`, `#cork`, `#ireland`, `#island`, `#lighthouse`, `#ocean`, `#sea`

GT hashtags: `#alaska`, `#glacier`, `#juneau`

Figure 4. Image-to-hashtag retrieval. The user history image (left) show the word cloud representation of the user history, and the size of the words is proportional to the frequency of the words in the user history. Given the user representation and the image (center), our model predicts some hashtags (right). The size of the hashtags is proportional to the probability of the hashtags.

The result of Figure 5 is very interesting because it shows that pretrained word embeddings allows to deal with several languages. We show the predicted hashtags of an image for a user using several languages: English, French and Spanish. The left image shows the word cloud representation of the user history. We can see that the same word is used in several languages *e.g.* (*night* 🇬🇧 = *nuit* 🇫🇷, *light* 🇬🇧 = *lumière* 🇫🇷, *museum* 🇬🇧 = *musée* 🇫🇷). Given a new image, the model predicts the same hashtags in different languages because it does not which language to choose:

- *sea* 🇬🇧 = *mer* 🇫🇷 = *mare* 🇪🇸
- *beach* 🇬🇧 = *plage* 🇫🇷 = *playa* 🇪🇸
- *sand* 🇬🇧 = *sable* 🇫🇷
- *water* 🇬🇧 = *eau* 🇫🇷



USER HISTORY IMAGE PREDICTED HASHTAGS

Figure 5. Multi-language hashtag retrieval. The user history image (left) show the word cloud representation of the user history, and the size of the words is proportional to the frequency of the words in the user history. Given the user representation and the image (center), our model predicts some hashtags (right). We observe that the model is able to predict the same hashtag in different languages *e.g. sea* = *mer* = *mare*.

A.5. Hashtag-to-image retrieval

We consider the hashtag-to-image retrieval task: given a query hashtag, find images that match the hashtag. A key challenge in this task is that hashtags can have multiple meanings: ideally, retrieval methods retrieve images corresponding to all meanings of a hashtag. We measure the performances with the Precision@10, *i.e.*, the fraction of the 10 top-scoring images that have the query hashtag associated with it. We also evaluate the performances on the 1k most frequent hashtags, denoted P@10 (1000). Table 2 presents the hashtag-based image retrieval performances on the test set. We observe that our user representation is better than others approaches, and the conclusions are the same that for hashtag retrieval.

MODEL	P@10	P@10 (1000)
[B] user agnostic	0.55	10.85
[E] hashtag sum	16.81	59.51
Ours (hashtag)	19.54	77.30
Ours (image+hashtag)	**20.84**	**81.16**

Table 2. Hashtag-based image retrieval. We compare several strategy to extract a user representation based on user image history.

We also show retrieved images for several hashtags in Figure 6, 7, 8, 9 and 10: `#artwork`, `#australia`, `#basketball`, `#empire state building`, `#football`, `#fun`, `#happy`, `#lego`, `#nature`, `#paris`, `#rock`, `#starwars`, `#underwater`, `#water`, `#waterfall`. These hashtags represent a wide variety of visual concepts. We note that for each hashtag (excepts `#underwater`), the images are retrieved from several users. The reason is probably that the `#underwater` tag is very specific and taking a picture underwater requires special camera. We observe that the retrieved photos are relevant for their corresponding hashtag even if some images are not labeled with the hashtags, which can explain that the low Precision@10 performance in Table 2. Our model is able to retrieve images from hashtags with multiple meanings. For instance, the retrieved images for `#rock` contains photos of rock as stone and rock as music. Similarly, the retrieved images for `#football` contains photos of (American) football and football (soccer).

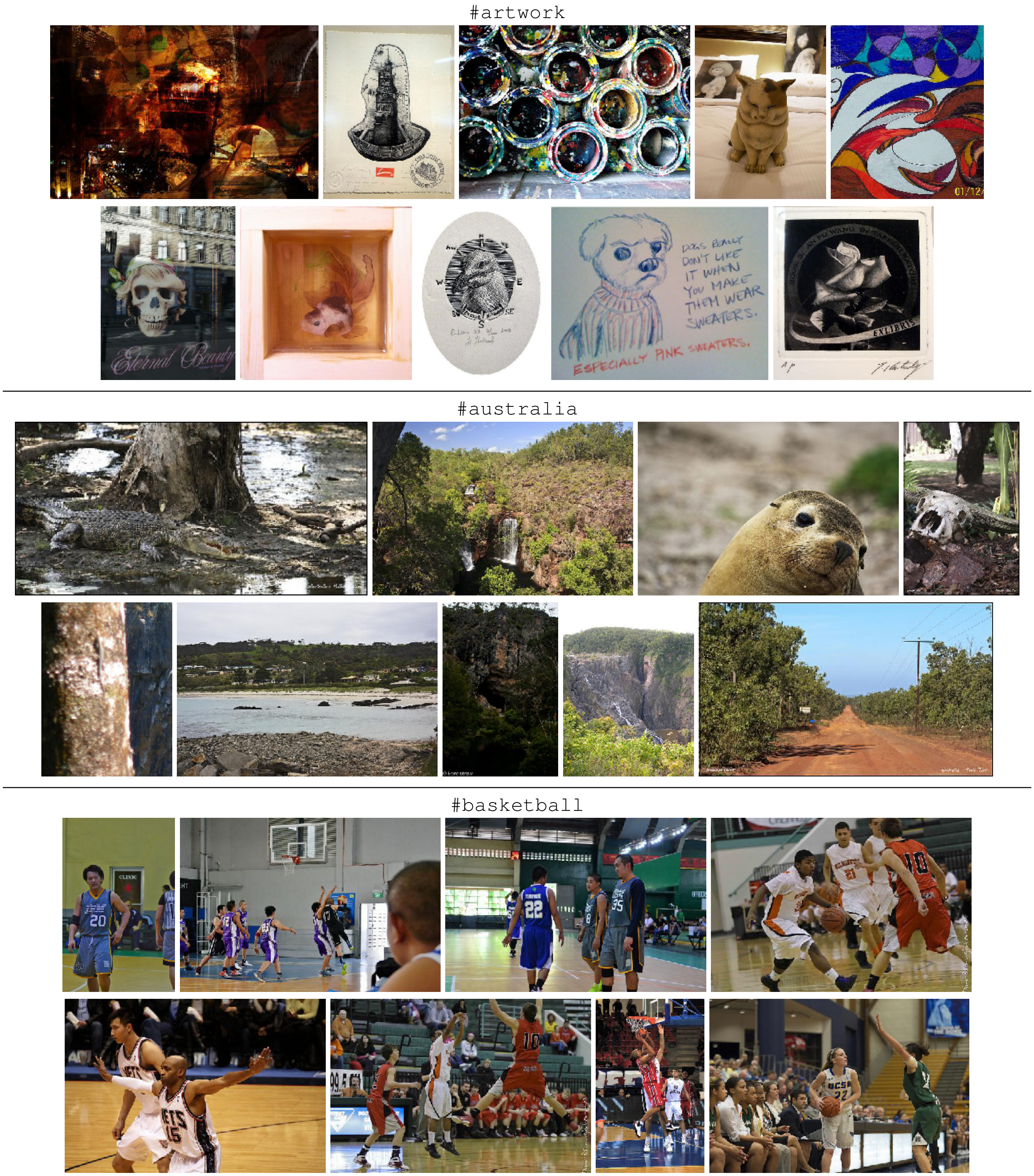
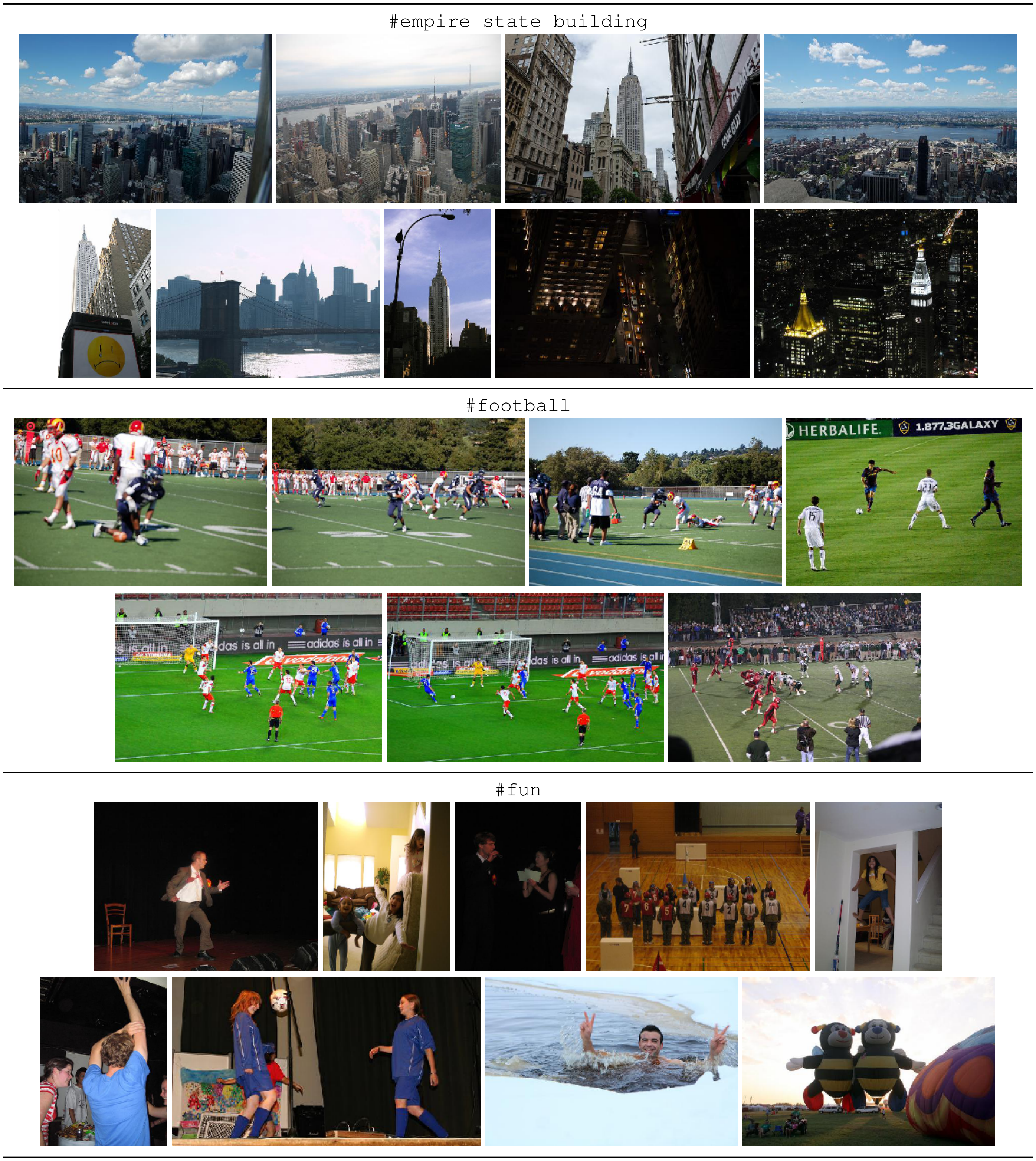

#artwork
01/12/
Eternal Beauty
DOGS REALLY DON'T LIKE IT WHEN YOU MAKE THEM WEAR SWEATERS.
ESPECIALLY PINK SWEATERS.
EXLIBRIS
#australia
#basketball

Figure 6. Hashtag-based image retrieval. Retrieved images with high probability for #artwork, #australia and #basketball.

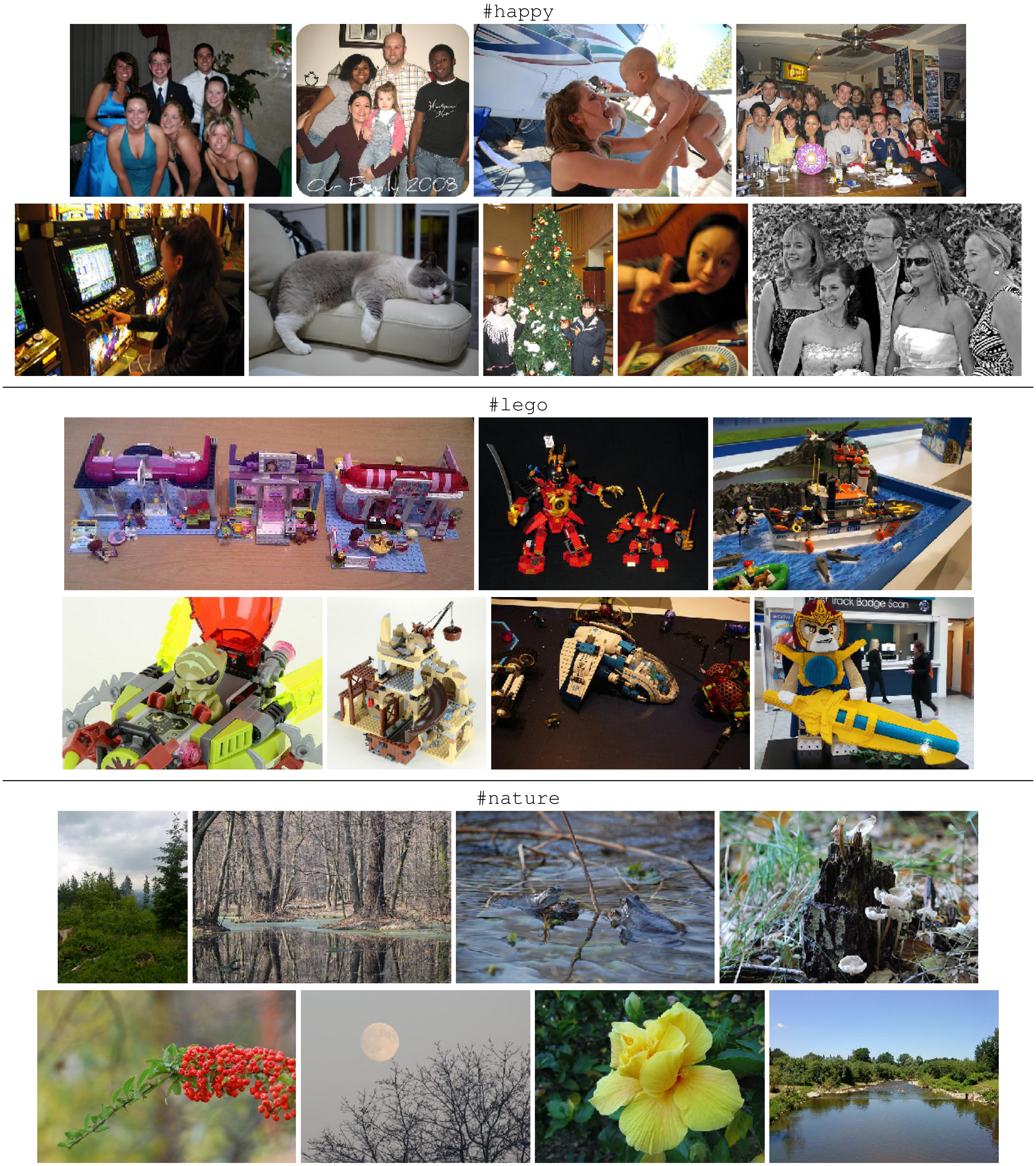


#empire state building

#football

#fun

Figure 7. Hashtag-based image retrieval. Retrieved images with high probability for `#empire state building`, `#football` and `#fun`.



#happy

#lego

#nature

Figure 8. Hashtag-based image retrieval. Retrieved images with high probability for #happy, #lego and #nature.

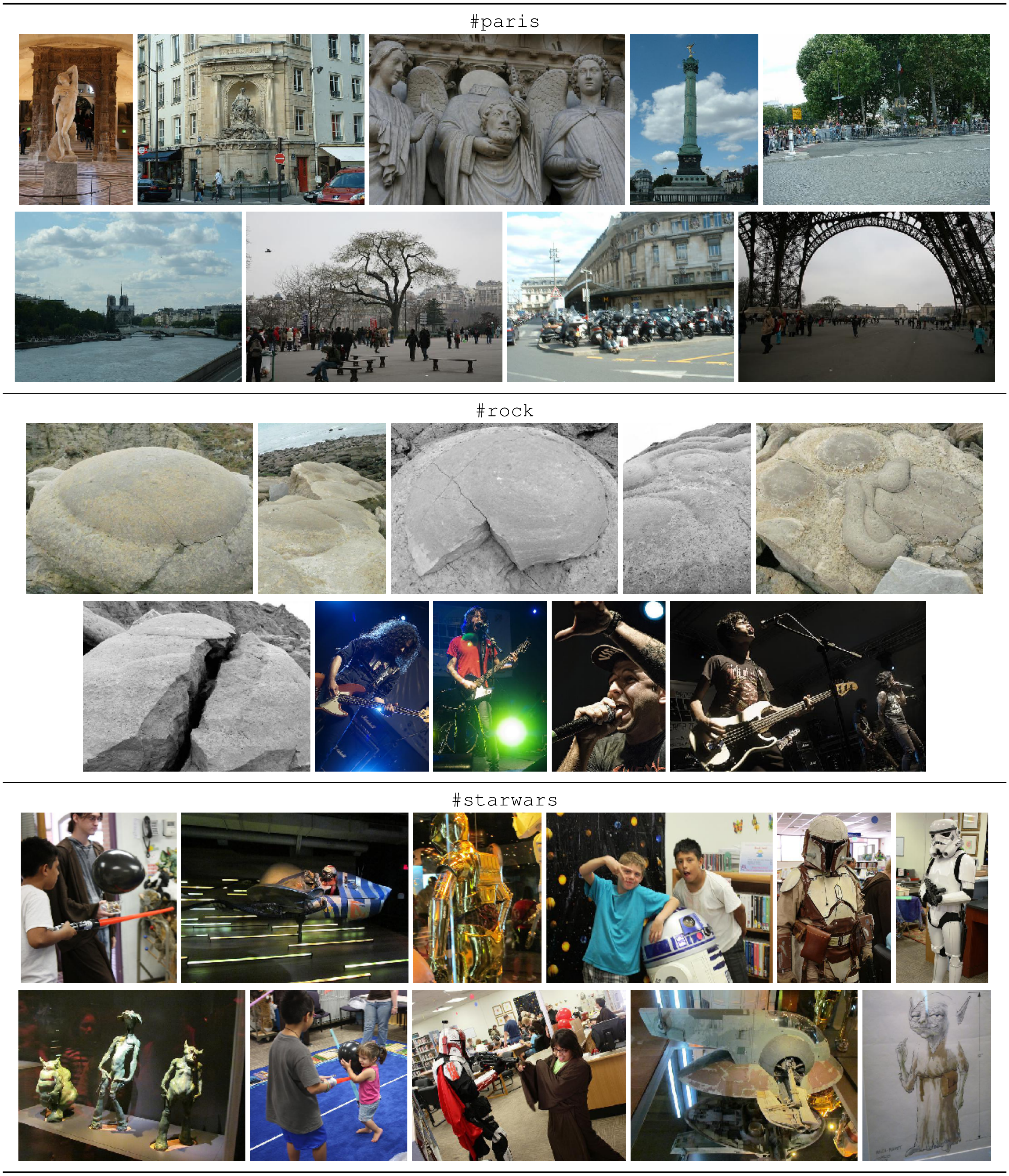


Figure 9. Hashtag-based image retrieval. Retrieved images with high probability for #rock, #paris and #starwars.

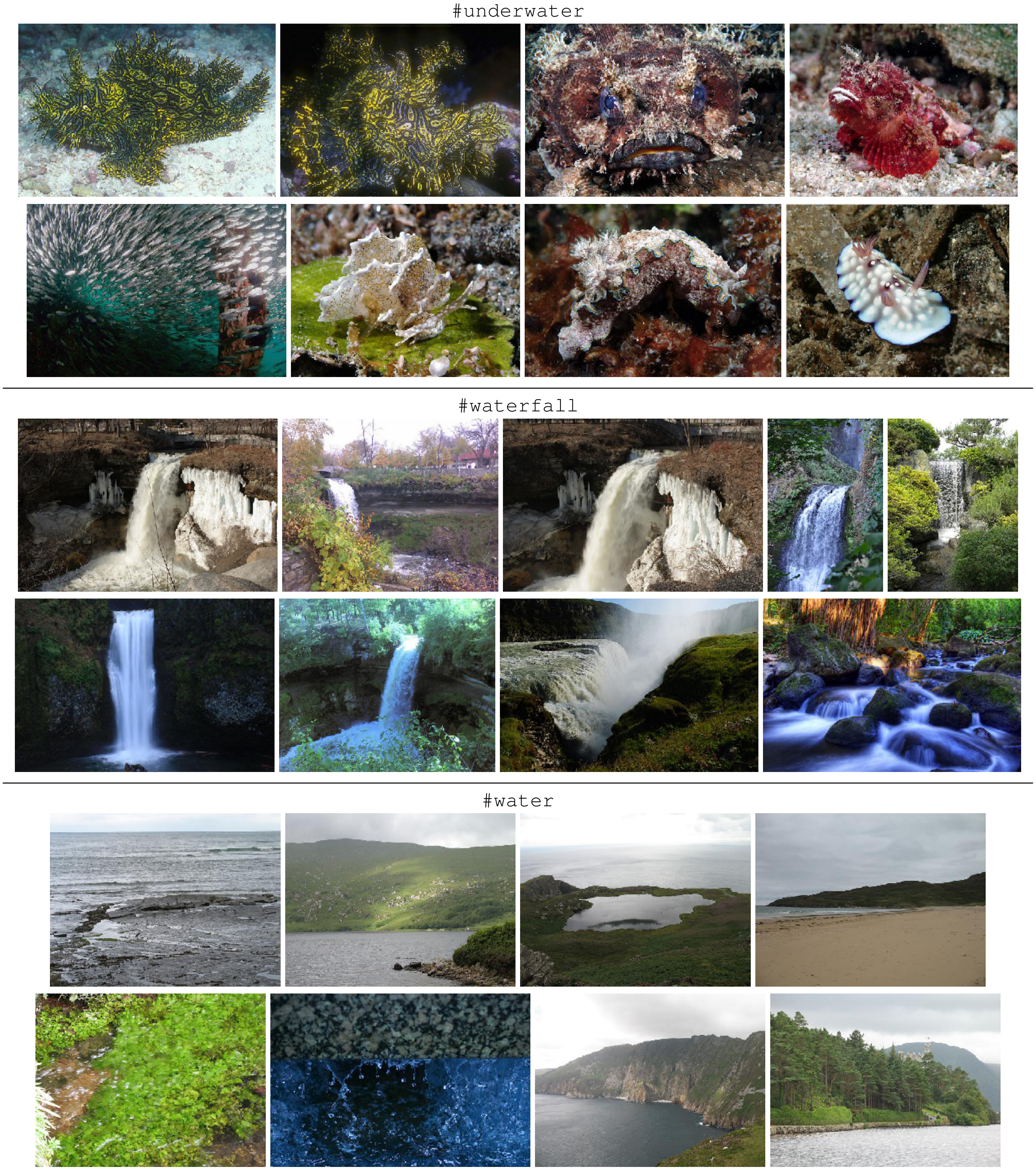


Figure 10. Hashtag-based image retrieval. Retrieved images with high probability for #underwater, #waterfall and #water.

A.6. Model analysis

A.6.1 Analysis of the history size

We analyze the importance of the history size *i.e.* the number of images used to compute the user representation. For each user, the first 50 images are used to build the user history and the remaining images are used for testing (users with less than 51 images are ignored). For instance for the history size of 10, we use the 40-th to the 49-th (included) images to compute the user representation. The results for different user history sizes are shown in Table 3. We observe that our model with only one image in the user history is significantly better than a user agnostic model. Our model can compute an accurate user representation with few images, and increasing the number of images in the user history improves the performance on all metrics.

This analysis shows that using one image and its asssociated hashtags give a lot of information about the user, like language (the dataset used in our experiments contains a lot of languages). Moreover, if there is a geographical hashtag like *Australia*, this hashtag also gives infomation about the location of the person and the next pictures will be probably about other Australian places.

HISTORY SIZE	A@1	A@10	P@10	R@1	R@10
0	7.55	21.04	3.60	2.27	6.93
1	29.94	54.76	17.93	10.44	27.68
2	30.84	54.87	19.13	11.05	28.68
5	31.67	55.55	19.72	11.41	29.35
10	32.28	56.33	19.97	11.58	29.77
20	32.68	56.85	20.15	11.70	30.01
30	32.81	57.02	20.20	11.72	30.09
40	32.90	57.17	20.24	11.75	30.16
50	**33.08**	**57.56**	**20.30**	**11.76**	**30.34**

Table 3. Analysis of the importance of the history size *i.e.* the number of images used to compute the user representation. The history size of 0 is a user agnostic model.

A.6.2 Image and hashtags complementarity

The hashtags are more important than the images in the user model because hashtags are inherently subjective whereas images are almost objective. Hashtags are provided by users as a source of self-expression. Unlike image, hashtags can provide non visual information *e.g.* an image does not give information about the language of the user. It is difficult to predict the hashtags of an image if we do not know the language of the user (the dataset used in our experiments contains more than 30 languages). We observe for all our experiments that using the image in the user model improves the performances which validates the fact that images also contain information about the user. The model can learn some patterns between the visual content of the images and the hashtags for each user.

We also observed that some geographical hashtags are difficult to predict. For instance, it is easy to predict the hashtag `#france` if it is a picture of the Eiffel Tower but it is difficult if it is a picture of a beach (Figure 5 of supplementary). However, if there are previously pictures of Paris, it becomes easier for the model to predict the location of the beach because the subset of possible answers is smaller. Hashtags about feeling e.g. `#happy` are also difficult to predict because they depend of the user and do not have clear visual appearance.

References

[1] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating Self-Expression and Visual Content in Hashtag Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3