Location Sensitive Image Retrieval and Tagging Supplementary Material

Raul Gomez[1,2], Jaume Gibert[1], Lluis Gomez[2], and Dimosthenis Karatzas[2]

Eurecat, Centre Tecnològic de Catalunya, Unitat de Tecnologies Audiovisuals
Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
{raul.gomez,jaume.gibert}@eurecat.org {lgomez,dimos}@cvc.uab.es

1 LocSens vs Visual Agnostic Tagging

LocSens has the capability of jointly modeling visual and location information to assign better contextualized tags, and inferred tags are generally related to the image content while clearly conditioned by the image location, as shown in paper's Figures 3 and 4. However, image location by itself is a powerful information to infer tags, since the words with which users tag their images are highly dependent on location. In fact, in addition to tags related with image content, images are usually tagged with the name of the place where they were taken. Images with places names as tags are particularly common in the YFCC100M dataset used in this research, since most of the images are from photographer's travels which tend to tag their uploaded images with their travels destinations. In this section we quantify how useful location information is if it is not jointly interpreted with visual information, and compare unimodal tagging performance with *LocSens* performance. We then show how *LocSens* goes beyond predicting places names, jointly interpreting visual and location information to assign better contextualized tags related to the image content.

1.1 Location Based Baselines

YFCC100M dataset provides also country, region and town names associated with each image, which have been specified by the user or inferred from the location. We computed the most frequent tags for each country and town in the training set. Then, we tagged each test image with the most common tags in its location to evaluate visual agnostic location based tagging baselines. Table 1 shows the performance of these baselines, the Multi-Class Classification location agnostic model and *LocSens*. Location based baselines scores are high, and the *Town Frequency* baseline outperforms the MCC (the best location agnostic baseline) in all metrics. It also outperforms *LocSens* in A@1 and reaches a close score in A@10. However *LocSens* A@50 score is superior by a large margin to unimodal models. There are two reasons why location based baselines show high performances:

Most of the YFCC100M images (78%) are tagged with places names. Places names are actually among the most common tags in the dataset. For instance:
Top global tags: london, unitedstates, england, nature, europe, japan, art, music, newyork, beach
Top United States tags: unitedstates, newyork, sanfrancisco, nyc, washington, texas, florida, chicago, seattle
Top San Francisco tags: sanfrancisco, sf, unitedstates, francisco, san, iphone, protest, gay, mission

In the A@*k* metric it is enough to correctly infer one image tag to get the maximum score for that image. Therefore, since most of the images are tagged with places names, a tagging method solely based on location that does not predict tags related to the image content can get high scores. As an example, if an image is tagged with *sydney*, *beach*, *sand* and *dog*, a method predicting only *sydney* from those tags would get the same A@k score as a method predicting all of them. However, we use A@k because is a standard performance metric for tagging and because it is also adequate to evaluate how *LocSens* exploits location to outperform location agnostic models.

LocSens outperforms the location baselines in A@50 by a big margin. One of the reasons is that *LocSens* is also predicting correct tags for those images that do not have places names as tags.

Table 1. Image tagging: Accuracy@1, accuracy@10 and accuracy@50 of two visual agnostic hashtag prediction models, MCC and the location sensitive model.

Method	*A*@1	*A*@10	*A*@50
Country Frequency	28.05	46.63	63.14
Town Frequency	**51.41**	65.49	71.05
MCC	20.32	47.64	68.05
LocSens - Raw locations	28.10	**68.21**	**85.85**

1.2 Beyond Places Names

Location based baselines achieve high A@k scores by predicting places names as tags because most of the images are tagged with them. However, *LocSens*, besides predicting tags related to image content and tags directly related to the given location, it predicts tags given the joint interpretation of visual and location information. To evaluate this behaviour, we omitted places names from groundtruth, frequency baselines and inferences and evaluated the methods. We construct the places list to omit by gathering all the continents, countries, regions and towns names in YFCC100M. Table 2 shows the results. All performances

are significantly worse, which is due to the less amount of groundtruth tags. *LocSens* performs much better than the best location agnostic model (MCC) even in this setup, where predicting places names tags is not evaluated. This proves that *LocSens* goes beyond that, exploiting location information to jointly interpret visual and location information to predict better contextualized tags. In this case, *LocSens* performs also much better than the location based baselines, since the reason of their high performance is their accuracy predicting places names, as explained in the former section.

Table 2. Image tagging omitting places names: Accuracy@1, accuracy@10 and accuracy@50 of two visual agnostic hashtag prediction models, MCC and the location sensitive model.

Method	*A*@1	*A*@10	*A*@50
Country Frequency	3.80	17.21	41.60
Town Frequency	16.97	34.95	47.53
MCC	15.15	36.75	51.80
LocSens - Raw locations	**17.34**	**44.45**	**61.10**

2 Results Analysis

2.1 Retrieval

Beyond retrieving common images at each location. Paper's Figure 1 and Figure 1 in this supplementary material show *LocSens* retrieval results for the hashtag *temple* and *bridge* at different locations. They demonstrate how *LocSens* is able to distinguish between images related to the same concept across a wide range of cities with different geographical distances between them. Note that, despite some specific bridges might have a huge amount of images tagged with *bridge* in the dataset, as the San Francisco bridge or the Brooklyn bridge in New York, the system manages to retrieve images of other less represented bridges around the world. So, first and despite the bridges samples unbalance, it is learning to extract visual patterns that generalize to many different bridges around the world and, second, it is correctly balancing the tag query and location query influence in the final score. Paper's Figure 5 shows *LocSens* results for hashtags queries in different locations. The model is able to retrieve images related to a wide range of tags, from tags referring to objects, such as *car*, to tags referring to more abstract concepts, such as *hiking*, from the 100.000 tags vocabulary. It goes beyond learning the most common images from each geographical location, as it is demonstrated by the *hiking* results in El Cairo or the *car* results in Paris, which are concepts that do not prevail in images in those locations, but the system is still able to accurately retrieve them.

Challenging queries. Figure 2 shows *LocSens* results for hashtag queries in different locations where some queries are incompatible because the hashtag refers to a concept which does not occur in the given location. When querying with the *beach* hashtag in a coastal location such as Auckland, *LocSens* retrieves images of close-by beaches. But when we query for *beach* images from Madrid, which is far away from the coast, we get bullfighting and beach volley images, because the sand of both arenas makes them visually similar to beach images. If we try to retrieve beach images near Moscow, we get scenes of people sunbathing. Similarly, if we query for *ski* images in El Cairo and Sydney, we get images of the dessert and water sports respectively, which have visual similarities with ski images.

P@10 depending on hashtag frequency. Figure 3 shows the P@10 score on location agnostic image retrieval for the MLC, the MCC and the HER training methods for query tags as a function of their number of appearances on the training set. It shows that all methods perform better for query hashtags that are more frequent in the training data, but MCC significantly outperforms the other methods also in less frequent hashtags.

P@10 per continent at country granularity. Figure 4 shows the number of training images per continent, and the P@10 at country level (750 km) per continent of the *LocSens* model performing better at it ($\sigma = 0$). It shows how, with the exception of the Asia, the precision at country level is higher for continents with a bigger amount of training images.

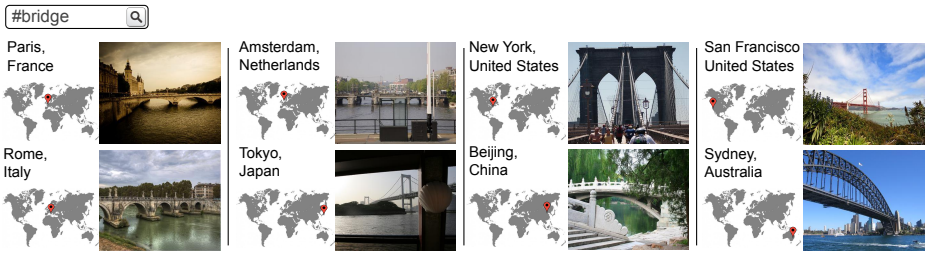

Fig. 1. Top retrieved image by the location sensitive model for the query hashtag "*bridges*" at different locations.

2.2 Tagging

Paper's Figure 4 and Figure 5 of this supplementary material show *LocSens* tagging results for images with different faked locations. They demonstrate that *LocSens* is able to exploit locations to assign better contextualized tags, jointly

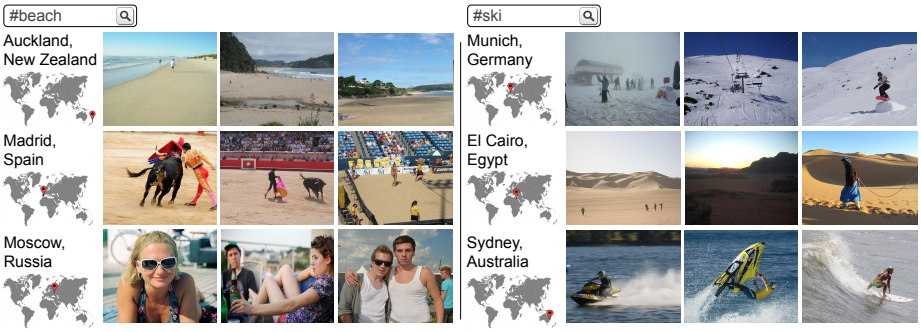

Fig. 2. Query hashtags with different locations where some queries are incompatible because the hashtag refers to a concept which does not occur in the query location.

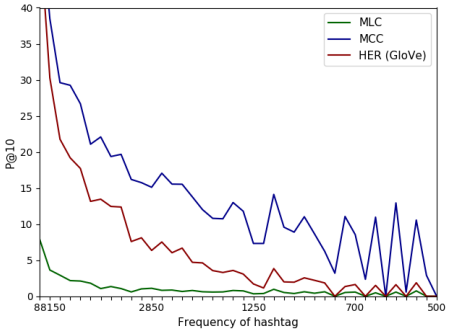

Fig. 3. Image Retrieval P@10 per hashtag as a function of the number of hashtag appearances in the training set for the MLC, the MCC and the HER models.

interpreting both query visual and location modalities. For instance, it assigns to the river image *lake* and *westlake* if it is from Los Angeles, since Westlake is the nearest important water geographic accident, while if the image is from Rio de Janeiro it tags it with *amazonia* and *rainforest*, and with *nile* if it is from El Cairo. In the example of an image of a road, it predicts as one of the most probable tags *carretera* (which means *road* in spanish) if the image is from Costa Rica, while it predicts *hills*, *Cumbria* and *Scotland* if the image is from Edinburgh, referring to the geography and the regions names around. If the image is from Chicago, it predicts *interstate*, since the road in it may be from the United States interstate highway system. These examples prove the joint interpretation of the visual and the location modalities to infer the most probable tags, since predicted tags are generally related to the image content while clearly conditioned by the image location.

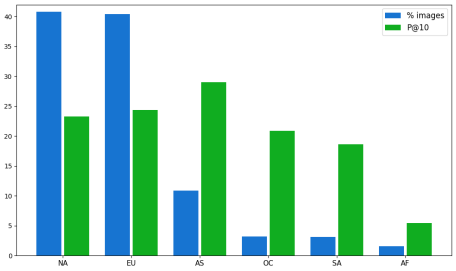

Fig. 4. Number of training images per continent and Location Sensitive Image Retrieval P@10 at country granularity (750 km) per continent.

Los Angeles, United States	Rio Jaineiro, Brazil	El Cairo, Egypt
#water	#river	#egypt
#lake	#amazonia	#river
#westlake	#paraty	#water
#river	#rio	#nile
#reflection	#rainforest	#nature

Chicago, UUEE	San José, Costa Rica	Edinburgh, UK
#road	#road	#road
#highway	#highway	#hills
#ohio	#rural	#sign
#interstate	#centralamerica	#cumbria
#sign	#carretera	#scotland

Fig. 5. *LocSens* top 5 predicted hashtags for images with 3 different faked locations.

3 Location Relevance in Image Retrieval

The reason why the P@10 score difference between MCC and *LocSens* on location sensitive image retrieval (shown in Table 1) is small is because the location information is not useful for many queries in our set because of their hashtags. There are several reasons for which a query hashtag can make the query location conditioning useless:

- **Hashtags carrying explicit location information.** Query hashtags that carry explicit location information are numerous in our query set, given it contains many travel pictures (i.e New York, Himalaya, Amazonas). See most frequent tags in the first section of this supplementary material.
- **Hashtags carrying implicit location information.** Query hashtags that do not refer to specific locations, but carry implicit information of it. For instance, the language of the hashtag can indicate its location. Also hashtags referring to local celebrations, local dishes, etc.
- **Hashtags with a visual appearance invariant to location**. Query hashtags that have the same visual appearance worldwide (such as "cat" or "tomato"), for which location-specific image features cannot be learnt.

Therefore, the performance improvements of LocSens compared to MLC reported on Table 1 are small because location is irrelevant in many queries of this particular dataset, so LocSens is only able to outperform MLC in a small percentage of them. Besides, although MCC and *LocSens* P@10 might be close, are qualitatively different an they do no retrieve the same images: As an example, LocSens - Raw locations retrieves images that are always near to the query

location, but gets worse results than MCC in continent and country granularities because their relation with the query tag is weaker. In this work we have focused on learning from large scale Social Media data. Further experimentation under more controlled scenarios where the location information is meaningful in all cases is another interesting research setup to evaluate the same tasks.

4 Implementation Details

4.1 MLC

The training of the MLC model was very unstable because of the class imbalance. We did try different class-balancing techniques without consistent improvements, and concluded that it is not an adequate training setup for our problem. We also tried different methods to evaluate both image tagging and retrieval using the MLC method, such as directly ranking the tags or the images with the scores, or learning embeddings with an intermediate 300-d layer as we do with MCC, but all experiments led to poor results.

4.2 LocSens

LocSens is trained with precomputed images and tag embeddings to reduce the computational load. Also, given LocSens has as inputs images but also tags embeddings learned by MCC, an architecture jointly optimizable would not be straight forward.

LocSens maps the image, tag, and location modalities to 300-d representations and then concatenates them. We experimented merging 2-d locations with the other modalities but couldn't optimize *LocSens* properly. We tried different strategies such as initializing LocSens parameters to attend location values, but mapping the three modalities to the same dimensionality before their concatenation yielded the best results. This is probably because it allows the model to better balance the different modalities.

To train *LocSens* with the location sampling technique we start always form $\sigma = 1$ and slowly decrease it to get models sensitive to different location granularities, evaluated in Table 1..

5 Future Work

The presented work can give rise to further research on how to exploit location information in image retrieval and tagging tasks, and also on how to learn image representations with tags supervision from large scale weakly annotated data. We spot three different experimentation lines to continue with this research work:

– **Learning with tags supervision.** Our research on learning image representations with hashtags supervision concludes that a Multi-Class setup with Softmax activations and a Cross-Entropy loss outperforms the other

baselines by a big margin. A research line to uncover the reason for this superior performance and to find under which conditions this method outperforms other standard learning setups, such as using a Multi-Label setup with Sigmoid activations, would be very interesting for the community.

- **More efficient architectures.** The current efficiency of the method is a drawback, since for instance to find the top tags for an image and location query, we have to compute the score of the query with all the hashtags in the vocabulary. An interesting research line is to find architectures for the same task that are more efficient than LocSens. As an example, we have been researching on tagging models that learn a joint embedding space for hashtags and image+location pairs, which at inference time only need to compute a distance between an image+location query embedding and precomputed tags embeddings, being much more efficient. The drawback of such architectures is, however, that the same model cannot be used for tagging and retrieval as LocSens can: A retrieval model with this architecture would have to learn a joint embedding space for hashtags+location pairs and images.
- **Information modalities balance.** In the paper we propose a location sampling strategy useful to balance the location influence in the image ranking. Experimentation on how this technique can be exploited in other multimodal tasks would be an interesting research line.