Location Sensitive Image Retrieval and Tagging

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

Abstract. People from different parts of the globe describe objects and concepts in distinct manners. Visual appearance can thus vary across different geographic locations, which makes location a relevant contextual information when analysing visual data. In this work, we address the task of image retrieval related to a given tag conditioned on a certain location on Earth. We present LocSens, a model that learns to rank triplets of images, tags and coordinates by plausibility, and two training strategies to balance the location influence in the final ranking. LocSens learns to fuse textual and location information of multimodal queries to retrieve related images at different levels of location granularity, and successfully utilizes location information to improve image tagging.

1 Introduction

Image tagging is the task of assigning tags to images, referring to words that describe the image content or context. An image of a beach, for instance, could be tagged with the words *beach* or *sand*, but also with the words *swim*, *vacation* or *Hawaii*, which do not refer to objects in the scene. On the other hand, image-by-text retrieval is the task of searching for images related to a given textual query. Similarly to the tagging task, the query words can refer to explicit scene content or to other image semantics. In this work we address the specific retrieval case when the query text is a single word (a tag).

Besides text and images, location is a data modality widely present in contemporary data collections. Many cameras and mobile phones with built-in GPS systems store the location information in the corresponding *Exif* metadata header when a picture is taken. Moreover, most of the web and social media platforms add this information to generated content or use it in their offered services. In this work we leverage this third data modality: using location information can be useful in an image tagging task since location-related tagging can provide better contextual results. For instance, an image of a skier in France could have the tags *"ski, alps, les2alpes, neige"*, while an image of a skier in Canada could have the tags *"ski, montremblant, canada, snow"*. More importantly, location can also be very useful in an image retrieval setup where we want to find images related to a word in a specific location: the retrieved images related to the query tag *temple* in Italy should be different from those in China. In this sense, it could be interesting to explore which kind of scenes people from different countries and

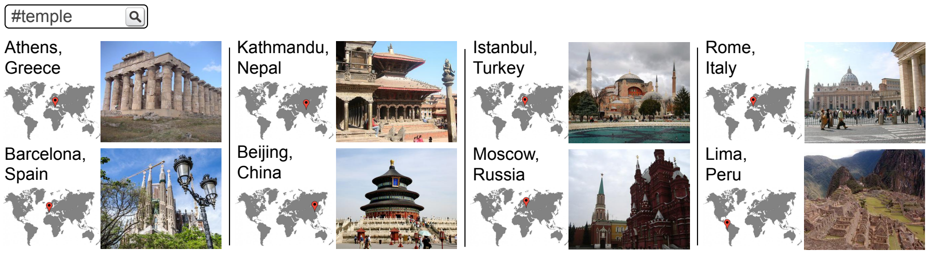

Fig. 1. Top retrieved image by *LocSens*, our location sensitive model, for the query hashtag "*temple*" at different locations.

cultures relate with certain *broader* concepts. Location sensitive retrieval results produced by the proposed system are shown in Figure 1.

In this paper we propose a new architecture for modeling the joint distribution of images, hashtags, and geographic locations and demonstrate its ability to retrieve relevant images given a query composed by a hashtag and a location. In this task, which we call location sensitive tag-based image retrieval, a retrieved image is considered relevant if the query hashtag is within its groundtruth hashtags and the distance between its location and the query location is smaller than a given threshold. Notice that distinct from previous work on GPS-aware landmark recognition or GPS-Constrained database search [13, 14, 24, 27] in the proposed task the locations of the test set images are not available at inference time, thus simple location filtering is not an option.

A common approach to address these situations in both image by text retrieval and image tagging setups is to learn a joint embedding space for images and words [6, 15, 28, 37]. In such a space, images are embedded near to the words with which they share semantics. Consequently, semantically *similar* images are also embedded together. Usually, word embedding models, such as Word2Vec [21] or GloVe [25] are employed to generate word representations, while a CNN is trained to embed images in the same space, learning optimal compact representations for them. Word models have an interesting and powerful feature: words with similar semantics have also similar representations and this is a feature that image tagging and retrieval models aim to incorporate, since learning a joint image and word embedding space with semantic structure provides a more flexible and less prone to drastic errors tagging or search engine.

Another approach to handle multiple modalities of data is by scoring tuples of multimodal samples aiming to get high scores on positive cases and low scores on negative ones [12, 29, 34, 38]. This setup is convenient for learning from Web and Social Media data because, instead of strict similarities between modalities, the model learns more relaxed compatibility scores between them. Our work fits under this paradigm. Specifically, we train a model that produces scores for image-hashtag-coordinates triplets, and we use these scores in a ranking loss in order to learn parameters that discriminate between observed and unobserved triplets. Such scores are used to tag and retrieve images in a location aware configuration providing good quality results under the large-scale YFCC100M dataset [33]. Our summarized contributions are:

- We introduce the task of location sensitive tag-based image retrieval.
- We evaluate different baselines for learning image representations with hashtag supervision exploiting large-scale social media data that serve as initialization of the location sensitive model.
- We present the *LocSens* model to score images, tags and location triplets (Figure 2), which allows to perform location sensitive image retrieval and outperforms location agnostic models in image tagging.
- We introduce novel training strategies to improve the location sensitive retrieval performance of *LocSens* and demonstrate that they are crucial in order to learn good representations of joint hashtag+location queries.

2 Related Work

Location-aware image search and tagging. O'Hare *et al.* [24] presented the need of conditioning image retrieval to location information, and targeted it by using location to filter out distant photos and then performing a visual search for ranking. Similar location-based filtering strategies have been also used for landmark identification [1] and to speed-up loop closure in visual SLAM [16]. The obvious limitation of such systems compared to *LocSens* is that they require geolocation annotations in the entire retrieval set. Kennedy *et al.* [13, 14] and Rattenbury *et al.* [27] used location-based clustering to get the most representative tags and images for each cluster, and presented limited image retrieval results for a subset of tags associated to a given location (landmark tags). They did not learn, however, location-dependent visual representations for tags as we do here, and their system is limited to the use of landmark tags as queries. On the other hand, Zhang *et al.* [47] proposed a location-aware method for image tagging and tag-based retrieval that first identifies points of interest, clustering images by their locations, and then represents the image-tag relations in each of the clusters with an individual image-tag matrix [42]. Their study is limited to datasets on single city scale and small number of tags (1000). Their retrieval method is constrained to use location to improve results for tags with location semantics, and cannot retrieve location-dependent results (i.e. only the tag is used as query). Again, contrary to *LocSens*, this method requires geolocation annotations over the entire retrieval set. Other existing location-aware tagging methods [17, 22] have also addressed constrained or small scale setups (e.g. a fixed number of cities) and small-size tag vocabularies, while in this paper we target a worldwide scale unconstrained scenario.

Location and Classification. The use of location information to improve image classification has also been previously explored. , and has recently experienced a growing interest by the computer vision research community. Yuan *et al.* [46] combine GPS traces and hand-crafted visual features for events classification. Tang *et al.* [32] propose different ways to get additional image context from coordinates, such as temperature or elevation, and test the usefulness of such information in image classification. Herranz *et al.* [10, 44] boost food dish classification using location information by jointly modeling dishes, restaurants and their menus and locations. Chu *et al.* [2] compare different methods to fuse visual and location information for fine-grained image classification. Mac *et al.* [18] also work on fine-grained classification by modeling the spatio-temporal distribution of a set of object categories and using it as a prior in the classification process. Location-aware classification methods that model the prior distribution of locations and object classes can also be used for tagging, but they can not perform location sensitive tag-based retrieval because the prior for a given query (tag+location) would be constant for the whole retrieval set.

Image geolocalization. Hays *et al.* [8] introduced the task of image geolocalization, i.e. assigning a location to an image, and used hand-crafted features to retrieve nearest neighbors in a reference database of geotagged images. Gallagher *et al.* [4] exploited user tags in addition to visual search to refine geolocalization. Vo *et al.* [35] employed a similar setup but using a CNN to learn image representations from raw pixels. Weyand *et al.* [39] formulated geolocalization as a classification problem where the earth is subdivided into geographical cells, GPS coordinates are mapped to these regions, and a CNN is trained to predict them from images. Müller-Budack *et al.* [23] enhanced the previous setup using earth partitions with different levels of granularity and incorporating explicit scene classification to the model. Although these methods address a different task, they are related to *LocSens* in that we also learn geolocation-dependent visual representations. Furthermore, inspired by [35], we evaluate our models' performance at different levels of geolocation granularity.

Multimodal Learning. Multimodal joint image and text embeddings is a very active research area. DeViSE [3] proposes a pipeline that, instead of learning to predict ImageNet classes, learns to infer the Word2Vec [21] representations of their labels. This work inspired others that applied similar pipelines to learn from paired visual and textual data in a weakly-supervised manner [6, 7, 30]. More related to our work, Veit *et al.* [34] also exploit the YFCC100M dataset [33] to learn joint embeddings of images and hashtags for image tagging and retrieval. They work on user-specific modeling, learning embeddings conditioned to users to perform user-specific image tagging and tag-based retrieval. Apart from learning joint embeddings for images and text, other works have addressed tasks that need the joint interpretation of both modalities. Although some recent works have proposed more complex strategies to fuse different data modalities [5, 20,26,36,45], their results show that their performance improvement compared to a simple feature concatenation followed by a Multi Layer Perceptron is marginal.

3 Methodology

Given a large set of images, tags and geographical coordinates, our objective is to train a model to score triplets of image-hashtag-coordinates and rank them to perform two tasks: (1) image retrieval querying with a hashtag and a location, and (2) image tagging when both the image and the location are available. We address the problem in two stages: first, we train a location-agnostic CNN to learn image representations using hashtags as weak supervision. We propose different training methodologies and evaluate their performance on image tagging and retrieval. These serve as benchmark and provide compact image representations to be later used within the location sensitive models. Second, using the learnt image and hashtags best performing representations and the locations, we train multimodal models to score triplets of these three modalities. We finally evaluate them on image retrieval and tagging and analyze how these models benefit from the location information.

3.1 Learning with hashtag supervision

Three procedures for training location-agnostic visual recognition models using hashtag supervision are considered: (1) multi-label classification, (2) softmax multi-class classification, and (3) hashtag embedding regression. In the following, let $\mathbb{H}$ be the set of H considered hashtags. $\mathbf{I_x}$ will stand for a training image and $\mathbf{H_x} \subseteq \mathbb{H}$ for the set of its groundtruth hashtags. The image model $f(\cdot;\theta)$ used is a ResNet-50 [9] with parameters θ. The three approaches eventually produce a vector representation for an image $\mathbf{I_x}$, which we denote by $\mathbf{r_x}$. For a given hashtag $h^i \in \mathbb{H}$, its representation —denoted $\mathbf{v_i}$— is either learnt externally or jointly with those of the images.

Multi-Label Classification (MLC). We set the problem in its most natural form: as a standard MLC setup over H classes corresponding to the hashtags in the vocabulary $\mathbb{H}$. The last ResNet-50 layer is replaced by a linear layer with H outputs, and each one of the H binary classification problems is addressed with a cross-entropy loss with sigmoid activation. Let $\mathbf{y_x} = (y_x^1, \ldots, y_x^H)$ be the multi-hot vector encoding the groundtruth hashtags of $\mathbf{I_x}$ and $\mathbf{f_x} = \sigma(f(\mathbf{I_x};\theta))$, where σ is the element-wise sigmoid function. The loss for image $\mathbf{I_x}$ is written as:

$$L = -\tfrac{1}{H}\sum_{h=1}^{H}[\, y_x^h \log f_x^h + (1 - y_x^h)\log(1 - f_x^h)\,]. \tag{1}$$

Multi-Class Classification (MCC). Despite being counter-intuitive, several prior studies [19, 34] demonstrate the effectiveness of formulating multi-label problems with large numbers of classes as multi-class problems. At training time a random target class from the groundtruth set $\mathbf{H_x}$ is selected, and softmax activation with a cross-entropy loss is used. This setup is commonly known as softmax classification.

Let $h_x^i \in \mathbf{H_x}$ be a randomly selected class (hashtag) for $\mathbf{I_x}$. Let also f_x^i be the coordinate of $\mathbf{f_x} = f(\mathbf{I_x};\theta)$ corresponding to h_x^i. The loss for image $\mathbf{I_x}$ is set to be:

$$L = -\log\left(\frac{e^{f_x^i}}{\sum_{j=1}^{H} e^{f_x^j}}\right). \tag{2}$$

In this setup we redefine ResNet-50 by adding a linear layer with D outputs just before the last classification layer with H outputs. This allows getting compact image D-dimensional representations $\mathbf{r_x}$ as their activations in such layer. Since we are in a multi-class setup where the groundtruth is a one-hot vector, we are also implicitly learning hashtag embeddings: the weights of the last classification layer with input $\mathbf{r_x}$ and output $\mathbf{f_x}$ is an $H \times D$ matrix whose rows can be understood as D-dimensional representations of the hashtags in $\mathbb{H}$. Consequently, this approach learns at once D-dimensional embeddings for both images and hashtags. In our experiments, the dimensionality is set to $D = 300$ to match that of the word embeddings used in the next and last approach. This procedure does not apply to MLC for which groundtruth is multi-hot encoded.

Hashtag Embedding Regression (HER). We use pretrained GloVe [25] embeddings for hashtags, which are D-dimensional with $D = 300$. For each image $\mathbf{I_x}$, we sum the GloVe embeddings of its groundtruth hashtags $\mathbf{H_x}$, which we denote as $\mathbf{t_x}$. Then we replace the last layer of the ResNet-50 by a D-dimensional linear layer, and we learn the parameters of the image model by minimizing a cosine embedding loss. If, $\mathbf{f_x} = f(\mathbf{I_x}; \theta)$ is the output of the vision model, the loss is defined by:

$$L = 1 - \left(\frac{\mathbf{t_x} \cdot \mathbf{f_x}}{\|\mathbf{t_x}\| \, \|\mathbf{f_x}\|} \right) . \tag{3}$$

As already stated by [34], because of the nature of the GloVe semantic space, this methodology has the potential advantage of not penalizing predicting hashtags with close meanings to those in the groundtruth but that a user might not have used in the image description. Moreover, as shown in [3] and due to the semantics structure of the embedding space, the resulting image model will be less prone to drastic errors.

3.2 Location Sensitive Model (*LocSens*)

We design a location sensitive model that learns to score triplets formed by an image, a hashtag and a location. We use a siamese-like architecture and a ranking loss to optimize the model to score positive triplets (existing in the training set) higher than negative triplets (which we create). Given an image $\mathbf{I_x}$, we get its embedding $\mathbf{r_x}$ computed by the image model, the embedding $\mathbf{v_{x_i}}$ of a random hashtag $h^i_{\mathbf{x}}$ from its groundtruth set $\mathbf{H_x}$ and its groundtruth latitude and longitude $\mathbf{g_x} = [\varphi_{\mathbf{x}}, \lambda_{\mathbf{x}}]$, which constitute a positive triplet. Both $\mathbf{r_x}$ and $\mathbf{v_{x_i}}$ are L2 normalized and latitude and longitude are both normalized to range in $[0, 1]$. Note that 0 and 1 latitude fall on the poles while 0 and 1 represent the same longitude because of its circular nature and falls on the Pacific.

The three modalities are then mapped by linear layers with ReLU activations to 300 dimensions each, and L2 normalized again. This normalization guarantees that the magnitudes of the representations of the different modalities are equal when processed by subsequent layers in the multimodal network. Then the three vectors are concatenated. Although sophisticated multimodal data fusion strategies have been proposed, simple feature concatenation has also been proven to be an effective technique [34,36]. We opted for a simple concatenation as it streamlines the strategy. The concatenated representations are then forwarded through 5 linear layers with normalization and ReLU activations with $2048, 2048, 2048, 1024, 512$ neurons respectively. At the end, a linear layer with a single output calculates the score of the triplet. We have experimentally found that Batch Normalization [11] hampers learning, producing highly irregular gradients. We conjecture that all GPU-allowable batch size is in fact a small batch size for the problem at hand, since the number of triplets is potentially massive and the batch statistics estimation will always be erratic across batches. Group normalization [43] is used instead, which is independent of the batch size and permits learning of the models.

To create a negative triplet, we randomly replace the image or the tag of the positive triplet. The image is replaced by a random one not associated with the tag $h^i_{\mathbf{x}}$, and the tag by a random one not in $\mathbf{H_x}$. We have found that the performance in image retrieval is significantly better when all negative triplets are created replacing the image. This is because the frequency of tags is preserved in both the positive and negative triplets, while in the tagging configuration less common tags are more frequently seen in negative triplets.

We train with a Margin Ranking loss, with a margin set empirically to $m = 0.1$, use 6 negative triplets per positive triplet averaging the loss over them, and a batch size of 1024. If s_x is the score of the positive triplet and s_n the score of the negative triplet, the loss is written as:

$$L = max(0, s_n - s_x + m). \tag{4}$$

Figure 2 shows the model architecture and also the training strategies to balance location influence, which are explained next.

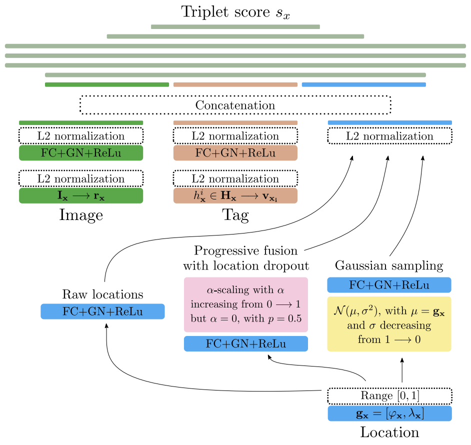
Triplet score s_x
Concatenation
L2 normalization
FC+GN+ReLu
L2 normalization
$\mathbf{I_x} \longrightarrow \mathbf{r_x}$
Image
L2 normalization
FC+GN+ReLu
L2 normalization
$h^i_{\mathbf{x}} \in \mathbf{H_x} \longrightarrow \mathbf{v}_{\mathbf{x}_i}$
Tag
L2 normalization
Raw locations
FC+GN+ReLu
Progressive fusion with location dropout
α-scaling with α increasing from $0 \longrightarrow 1$ but $\alpha = 0$, with $p = 0.5$
FC+GN+ReLu
Gaussian sampling
FC+GN+ReLu
$\mathcal{N}(\mu, \sigma^2)$, with $\mu = \mathbf{g_x}$ and σ decreasing from $1 \longrightarrow 0$
Range $[0, 1]$
$\mathbf{g_x} = [\varphi_{\mathbf{x}}, \lambda_{\mathbf{x}}]$
Location

Fig. 2. The proposed *LocSens* multimodal scoring model trained by triplet ranking (bars after concatenation indicate fully connected + group normalization + ReLu activation layers). During training, location information is processed and inputted to the model with different strategies.

Balancing Location Influence on Ranking. One important challenge in multimodal learning is balancing the influence of the different data modalities. We started by introducing the raw location values into the *LocSens* model, but immediately observed that the learning tends to use the location information to discriminate between triplets much more than the other two modalities, forgetting previously learnt relations between images and tags. This effect is especially severe in the image retrieval scenario, where the model ends up retrieving images close to the query locations but less related to the query tag. This suggests that the location information needs to be gradually incorporated into the scoring model for location sensitive image retrieval. For that, we propose the following two strategies, also depicted in Figure 2.

Progressive Fusion with Location Dropout. We first train a model with *LocSens* architecture but silencing the location modality hence forcing it to learn to discriminate triplets without using location information. To do that, we multiply by $\alpha = 0$ the location representation before its concatenation. Once the training has converged we start introducing locations progressively, by slowly increasing α until $\alpha = 1$. This strategy avoids new gradients caused by locations to ruin the image-hashtags relations *LocSens* has learned in the first training phase. In order to force the model to sustain the capability to discriminate between triplets without using location information we permanently zero the location representations with a 0.5 probability. We call this *location dropout* in a clear abuse of notation but because of its resemblance to zeroing random neurons in the well-known regularization strategy [31]. For the sake of comparison, we report results for the *LocSens* model with zeroed locations, which is in fact a location agnostic model.

Location Sampling. Exact locations are particularly narrow with respect to global coordinates and such a fine-grained degree of granularity makes learning troublesome. We propose to progressively present locations from rough precision to more accurate values while training advances. For each triplet, we randomly sample the training location coordinates at each iteration from a $2D$ normal distribution with mean at the image real coordinates ($\mu = \mathbf{g_x}$) and with standard deviation σ decreasing progressively. We constrain the sampling between [0, 1] by taking modulo 1 on the sampled values.

We start training with $\sigma = 1$, which makes the training locations indeed random and so not informative at all. At this stage, the *LocSens* model will learn to rank triplets without using the location information. Then, we progressively decrease σ, which makes the sampled coordinates be more accurate and useful for triplet discrimination. Note that σ has a direct relation with geographical distance, so location data is introduced during the training to be first only useful to discriminate between very distant triplets, and progressively between more fine-grained distances. Therefore, this strategy allows training models sensitive to different location levels of detail.

4 Experiments

We conduct experiments on the YFCC100M dataset [33] which contains nearly 100 million photos from Flickr with associated hashtags and GPS coordinates among other metadata. We create the hashtag vocabulary following [34]: we remove numerical hashtags and the 10 most frequent hashtags since they are not informative. The hashtag set $\mathbb{H}$ is defined as the set of the next 100,000 most frequent hashtags. Then we select photos with at least one hashtag from $\mathbb{H}$ from which we filter out photos with more than 15 hashtags. Finally, we remove photos without location information. This results in a dataset of 24.8M images, from which we separate a validation set of 250K and a test set of 500K. Images have an average of 4.25 hashtags.

4.1 Image by Tag Retrieval

We first study hashtag based image retrieval, which is the ability of our models to retrieve relevant images given a hashtag query. We define the set of querying hashtags $\mathbb{H}^q$ as the hashtags in $\mathbb{H}$ appearing at least 10 times in the testing set. The number of querying hashtags is $19,911$. If R_h^k is the set of top k ranked images for the hashtag $h \in \mathbb{H}^q$ and G_h is the set of images labeled with the hashtag h, we define precision@k as:

$$P@k = \frac{1}{|\mathbb{H}^q|} \sum_{h \in \mathbb{H}^q} \frac{|R_h^k \cap G_h|}{k}. \tag{5}$$

We evaluate precision@10, which measures the percentage of the 10 highest scoring images that have the query hashtag in their groundtruth. Under these settings, precision@k is upper-bounded by 100. The precision@10 of the different location agnostic methods described in Section 3.1 is as follows: MLC: 1.01, MCC: **14.07**, HER (GloVe): 7.02. The Multi-Class Classification (MCC) model has the best performance in the hashtag based image retrieval task.

4.2 Location Sensitive Image by Tag Retrieval

In this experiment we evaluate the ability of the models to retrieve relevant images given a query composed by a hashtag and a location (Figure 1). A retrieved image is considered relevant if the query hashtag is within its groundtruth hashtags and the distance between its location and the query location is smaller than a given threshold. Inspired by [35], we use different distance thresholds to evaluate the models' location precision at different levels of granularity. We define our query set of hashtag-location pairs by selecting the location and a random hashtag of $200,000$ images from the testing set. In this query set there will be repeated hashtags with different locations, and more frequent hashtags over all the dataset will also be more frequent in the query set (unlike in the location agnostic retrieval experiment of Section 4.1). This query set guarantees that the ability of the system to retrieve images related to the same hashtag but different locations is evaluated. To retrieve images for a given hashtag-location query with *LocSens*, we compute triplet plausibility scores with all test images and rank them.

Table 1 shows the performance of the different methods in location agnostic image retrieval and in different location sensitive levels of granularity. In location agnostic retrieval (first column) the geographic distance between the query and the results is not evaluated (infinite distance threshold). The evaluation in this scenario is the same as in Section 4.1, but the performances are higher because in this case the query sets contains more instances of the most frequent hashtags. The upper bound ranks the retrieval images containing the query hashtag by proximity to the query location, showcasing the optimal performance of any method in this evaluation. In location sensitive evaluations the optimal performance is less than 100% because we do not always have 10 or more relevant images in the test set.

Table 1. Location sensitive hashtag based image retrieval: P@10. A retrieved image is considered correct if its groundtruth hashtags contain the queried hashtag and the distance between its location and the queried one is smaller than a given threshold

		P@10					
	Method	**Location Agnostic**	**Continent (2500 km)**	**Country (750 km)**	**Region (200 km)**	**City (25 km)**	**Street (1 km)**
	Upper Bound	100	96.08	90.51	80.31	64.52	42.46
Img + Tag	MLC	5.28	2.54	1.65	1.00	0.62	0.17
	MCC	**42.18**	29.23	24.2	18.34	13.25	4.66
	HER (GloVe)	37.36	25.03	20.27	15.51	11.23	3.65
	LocSens - Zeroed locations	40.05	28.32	24.34	18.44	12.79	3.74
Loc + Img + Tag	LocSens - Raw locations	32.74	28.42	25.52	**21.83**	**15.53**	**4.83**
	LocSens - Dropout	36.95	30.42	26.14	20.46	14.28	4.64
	LocSens - Sampling $\sigma = 1$	40.60	28.40	23.84	18.16	13.04	4.13
	LocSens - Sampling $\sigma = 0.1$	40.03	29.30	24.36	18.83	13.46	4.22
	LocSens - Sampling $\sigma = 0.05$	39.80	31.25	25.76	19.58	13.78	4.30
	LocSens - Sampling $\sigma = 0.01$	37.05	**31.27**	26.65	20.14	14.15	4.44
	LocSens - Sampling $\sigma = 0$	35.95	30.61	**27.00**	21.39	14.75	**4.83**

Results show how the zeroed locations version of *LocSens* gets comparable results as MCC. By using raw locations in the *LocSens* model, we get the best results at fine level of location detail at the expense of a big drop in location agnostic retrieval. As introduced in Section 3.2, the reason is that it is relying heavily on locations to rank triplets decreasing its capability to predict relations between images and tags. As a result, it tends to retrieve images close to the query location, but less related to the query tag. The proposed dropout training strategy reduces the deterioration in location agnostic retrieval performance at a cost of a small drop in the fine levels of granularity. Also, it outperforms the former models in the coarse continent and country levels, due to its better balancing between using the query tag and location to retrieve related images. In its turn, the location sampling proposed approach with $\sigma = 1$ gets similar

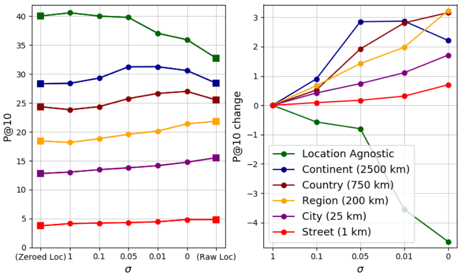

Fig. 3. Left: P@10 of the location sampling strategy for different σ and models with zeroed and raw locations. Right: P@10 difference respect to $\sigma = 1$.

results as *LocSens* with zeroed locations because the locations are as irrelevant in both cases. When σ is decreased, the model improves its location sensitive retrieval performance while maintaining a high location agnostic performance. This is achieved because informative locations are introduced to the model in a progressive way, from coarse to fine, and always maintaining triplets where the location is not informative, forcing the network to retain its capacity to rank triplets using only the image and the tag.

Figure 3 shows the absolute and relative performances at different levels of granularity while σ is decreased. At $\sigma = 0.05$, it can be seen that the location sensitive performances at all granularities have improved with a marginal drop on location agnostic performance. When σ is further decreased, performances at finer locations keep increasing, while the location agnostic performance decreases. When $\sigma = 0$, the training scenario is the same as in the raw locations one, but the training schedule allows this model to reduce the drop in location agnostic performance and at coarse levels of location granularity.

The location sampling technique provides *LocSens* with a better balancing between retrieving images related to the query tag and their location. Furthermore, given that σ has a direct geographical distance interpretation, it permits to tune the granularity to which we want our model to be sensitive. Note that *LocSens* enables to retrieve images related to a tag and near to a given location, which location agnostic models cannot do. The performance improvements in Table 1 at the different levels of location granularity are indeed significant since for many triplets the geographic location is not informative at all.

Figures 1 and 4 show qualitative retrieval results of several hashtags at different locations. They demonstrate that the model successfully fuses textual and location information to retrieve images related to the joint interpretation of the two query modalities, being able to retrieve images related to the same concept across a wide range of locations with different geographical distances between them. *LocSens* goes beyond retrieving the most common images from each geographical location, as it is demonstrated by the *winter* results in Berlin or the *car* results in Paris.

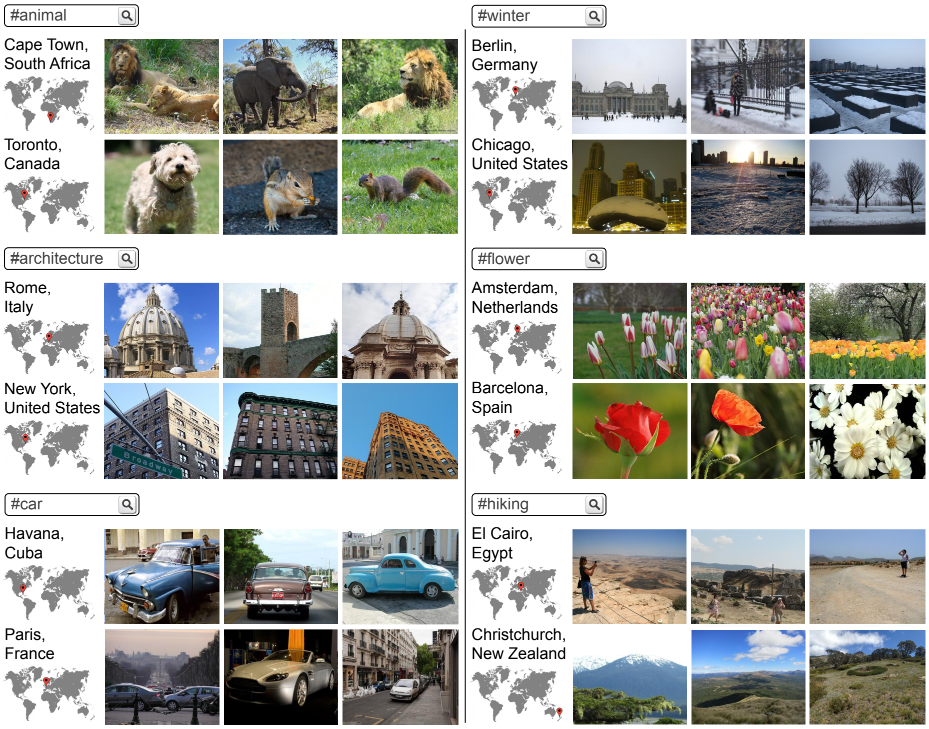

Fig. 4. Query hashtags with different locations and top 3 retrieved images.

4.3 Image Tagging

In this section we evaluate the ability of the models to predict hashtags for images in terms of $A@k$ (accuracy at k). If $\mathbf{H_x}$ is the set of groundtruth hashtags of $\mathbf{I_x}$, $\mathbf{R_x^k}$ denotes the k highest scoring hashtags for the image $\mathbf{I_x}$, and N is the number of testing images, $A@k$ is defined as:

$$A@k = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\mathbf{R_n^k} \cap \mathbf{H_n} \neq \emptyset\right], \tag{6}$$

where $\mathbb{1}[\cdot]$ is the indicator function having the value of 1 if the condition is fulfilled and 0 otherwise. We evaluate accuracy at $k = 1$ and $k = 10$, which measure how often the first ranked hashtag is in the groundtruth and how often at least one of the 10 highest ranked hashtags is in the groundtruth respectively.

A desired feature of a tagging system is the ability to infer diverse and distinct tags [40,41]. In order to measure the variety of tags predicted by the models, we measure the percentage of all the test tags predicted at least once in the whole test set (%`pred`) and the percentage of all the test tags correctly predicted at least once (%`cpred`), considering the top 10 tags predicted for each image.

Table 2 shows the performance of the different methods. Global Frequency ranks the tags according to the training dataset frequency. Among the location agnostic methods, MCC is the best one. This finding corroborates the experiments in [19, 34] verifying that this simple training strategy outperforms others when having a large number of classes. To train the *LocSens* model we used the image and tag representations inferred by the MCC model, since it is the one providing the best results.

Table 2. Image tagging: $A@1$, $A@10$, %`pred` and %`cpred` of the frequency baseline, location agnostic prediction and the location sensitive model

Method	$A@1$	$A@10$	%`pred`	%`cpred`
Global Frequency	1.82	13.45	0.01	0.01
MLC	8.86	30.59	8.04	4.5
MCC	**20.32**	**47.64**	**29.11**	**15.15**
HER (GloVe)	15.83	31.24	18.63	8.74
LocSens - Zeroed locations	15.92	46.60	26.98	13.31
LocSens - Raw locations	**28.10**	**68.21**	**44.00**	**24.04**

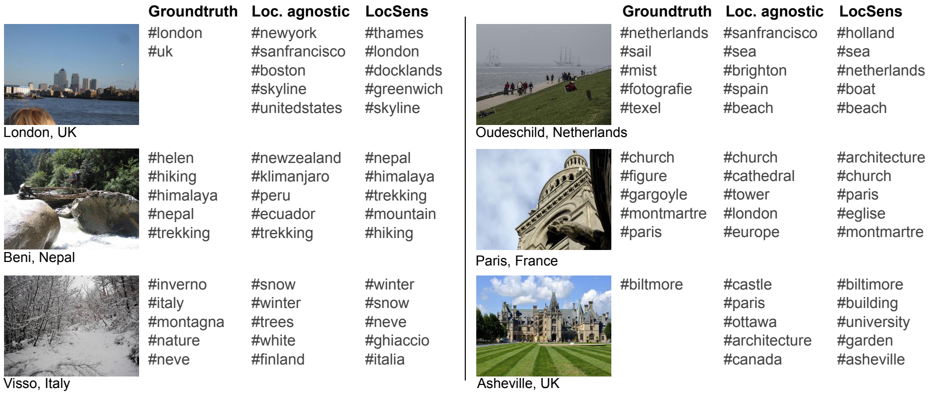

Fig. 5. Images with their locations and groundtruth hashtags and the corresponding top 5 hashtags predicted by the location agnostic MCC model and *LocSens*.

LocSens - Raw locations stands for the model where the raw triplets locations are always inputted both at train and test time. It outperforms the location agnostic methods in accuracy, successfully using location information to improve the tagging results. Moreover, it produces more diverse tags than location agnostic models, demonstrating that using location is effective for augmenting the hashtag prediction diversity. Figure 5 shows some tagging examples of a loca-

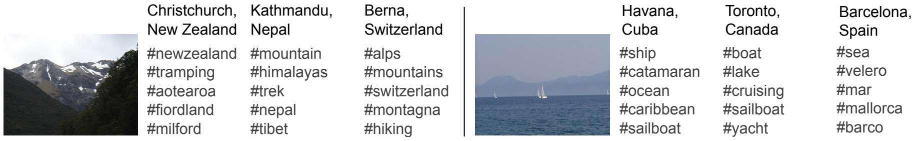

Fig. 6. *LocSens* top predicted hashtags for images with different faked locations.

tion agnostic model (MCC) compared to *LocSens*, that demonstrate how the later successfully processes jointly visual and location information to assign tags referring to the concurrence of both data modalities. As seen in the first example, besides assigning tags directly related to the given location (*london*) and discarding tags related to locations far from the given one (*newyork*), *LocSens* predicts tags that need the joint interpretation of visual and location information (*thames*). Figure 6 shows *LocSens* tagging results on images with different faked locations, and demonstrates that *LocSens* jointly interprets the image and the location to assign better contextualized tags, such as *caribbean* if a sailing image is from Cuba, and *lake* if it is from Toronto. Note that *LocSens* infers tags generally related to the image content while clearly conditioned by the image location, benefiting from the context given by both modalities. Tagging methods based solely on location, however, can be very precise predicting tags directly referring to a location, like place names, but cannot predict tags related to the image semantics. We consider the later a requirement of an image tagging system, and we provide additional experimentation in the supplementary material.

5 Conclusions

We have confirmed that a multiclass classification setup is the best method to learn image and tag representations when a large number of classes is available. Using them, we have trained *LocSens* to rank image-tag-coordinates triplets by plausibility. We have shown how it is able to perform image by tag retrieval conditioned to a given location by learning location-dependent visual representations, and have demonstrated how it successfully utilizes location information for image tagging, providing better contextual results. We have identified a problem in the multimodal setup, especially acute in the retrieval scenario: *LocSens* heavily relies on location for triplet ranking and tends to return images close to the query location and less related to the query tag. To address this issue we have proposed two novel training strategies: progressive fusion with location dropout, which allows training with a better balance between the modalities influence on the ranking, and location sampling, which results in a better overall performance and enables to tune the model at different levels of distance granularity.

Acknowledgement

Work supported by project TIN2017-89779-P, the CERCA Programme / Generalitat de Catalunya and the PhD scholarship AGAUR 2016-DI-84.

References

1. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., et al.: City-scale landmark identification on mobile devices. In: CVPR (2011)
2. Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., Leung, T., Adam, H.: Geo-Aware Networks for Fine-Grained Recognition. ICCVW (2019)
3. Frome, A., Corrado, G.S., Shlens, J., Bengio Jeffrey Dean, S., Ranzato, A., Mikolov, T.: DeViSE: A Deep Visual-Semantic Embedding Model. NIPS (2013)
4. Gallagher, A., Joshi, D., Yu, J., Luo, J.: Geo-location inference from image content and user tags (2009)
5. Gao, P., Lu, P., Li, H., Li, S., Li, Y., Hoi, S., Wang, X.: Question-Guided Hybrid Convolution for Visual Question Answering. ECCV (2018)
6. Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D., Jawahar, C.V.: Self-supervised learning of visual features through embedding images into text topic spaces. CVPR (2017)
7. Gordo, A., Larlus, D.: Beyond Instance-Level Image Retrieval: Leveraging Captions to Learn a Global Visual Representation for Semantic Retrieval. CVPR (2017)
8. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. CVPR (2008)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR (2016)
10. Herranz, L., Jiang, S., Xu, R.: Modeling Restaurant Context for Food Recognition. IEEE Transactions on Multimedia (2017)
11. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv (2015)
12. Jabri, A., Joulin, A., Van Der Maaten, L.: Revisiting Visual Question Answering Baselines. Tech. rep., Facebook AI Research (2016)
13. Kennedy, L., Naaman, M.: Generating Diverse and Representative Image Search Results for Landmarks. International Conference on World Wide Web (2008)
14. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. ACM International Conference on Multimedia (2007)
15. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. arXiv (2014)
16. Kumar, A., Tardif, J.P., Anati, R., Daniilidis, K.: Experiments on visual loop closing using vocabulary trees. In: CVPR Workshops (2008)
17. Liu, J., Li, Z., Tang, J., Jiang, Y., Lu, H.: Personalized geo-specific tag recommendation for photos on social websites. IEEE Transactions on Multimedia (2014)
18. Mac Aodha, O., Cole, E., Perona, P.: Presence-Only Geographical Priors for Fine-Grained Image Classification. ICCV (2019)
19. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der, L., Facebook, M.: Exploring the Limits of Weakly Supervised Pre-training. ECCV (2018)
20. Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic Multimodal Instance Segmentation guided by natural language queries. ECCV (2018)
21. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ICLR (2013)
22. Moxley, E., Kleban, J., Manjunath, B.: Spirittagger: a geo-aware tag suggestion tool mined from flickr. In: Proc. ACM ICMIR (2008)

23. Müller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: ECCV. vol. 11216 LNCS (2018)
24. O'Hare, N., Gurrin, C., Jones, G.J., Smeaton, A.F.: Combination of content analysis and context features for digital photograph retrieval. In: European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (2005)
25. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. EMNLP (2014)
26. Rajiv Jain, C.W.: Multimodal Document Image Classification. ICDAR (2019)
27. Rattenbury, T., Good, N., Naaman, M.: Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. ACM SIGIR Conference on Research and Development in Information Retrieval (2007)
28. Ren, Z., Jin, H., Lin, Z., Fang, C., Yuille, A.: Joint Image-Text Representation by Gaussian Visual-Semantic Embedding. ACM Multimedia (2016)
29. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of Textual Phrases in Images by Reconstruction. Tech. rep., Max Planck Institute for Informatics (2015)
30. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. CVPR (2017)
31. Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research (2014)
32. Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., Bourdev, L.: Improving Image Classification with Location Context. ICCV (2015)
33. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The New Data in Multimedia Research. Communications of the ACM (2015)
34. Veit, A., Nickel, M., Belongie, S., Maaten, L.V.D.: Separating Self-Expression and Visual Content in Hashtag Supervision. In: CVPR (2018)
35. Vo, N., Jacobs, N., Hays, J.: Revisiting IM2GPS in the Deep Learning Era. ICCV (2017)
36. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing Text and Image for Image Retrieval - An Empirical Odyssey. CVPR (2019)
37. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning Two-Branch Neural Networks for Image-Text Matching Tasks. CVPR (2017)
38. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. ICPR (2012)
39. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet-Photo Geolocation with Convolutional Neural Networks. ECCV (2016)
40. Wu, B., Chen, W., Sun, P., Liu, W., Ghanem, B., Lyu, S.: Tagging like Humans: Diverse and Distinct Image Annotation. CVPR (2018)
41. Wu, B., Jia, F., Liu, W., Ghanem, B.: Diverse Image Annotation. CVPR p. CVPR (2017)
42. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE TPAMI (2012)
43. Wu, Y., He, K.: Group Normalization. ECCV (2018)
44. Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., Jain, R.: Geolocalized Modeling for Dish Recognition. IEEE Transactions on Multimedia (2015)
45. Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., Predovic, G.: Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. Workshop on Abusive Language Online (2019)

46. Yuan, J., Luo, J., Kautz, H., Wu, Y.: Mining GPS Traces and Visual Words for Event Classification. Tech. rep., Northwestern University (2008)
47. Zhang, J., Wang, S., Huang, Q.: Location-Based Parallel Tag Completion for Geotagged Social Image Retrieval General Terms. ACM Transactions on Intelligent Systems and Technology (2017)