Learning Multi-context Aware Location Representations from Large-scale Geotagged Images

Yifang Yin
National University of Singapore
idsyin@nus.edu.sg

Ying Zhang
Northwestern Polytechnical University
yingz118@gmail.com

Zhenguang Liu
Zhejiang Gongshang University
liuzhenguang2008@gmail.com

Yuxuan Liang
National University of Singapore
yuxliang@outlook.com

Sheng Wang
Alibaba Group
Singapore
sh.wang@alibaba-inc.com

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Roger Zimmermann
National University of Singapore
rogerz@comp.nus.edu.sg

ABSTRACT

With the ubiquity of sensor-equipped smartphones, it is common to have multimedia documents uploaded to the Internet that have GPS coordinates associated with them. Utilizing such geotags as an additional feature is intuitively appealing for improving the performance of location-aware applications. However, raw GPS coordinates are fine-grained location indicators without any semantic information. Existing methods on geotag semantic encoding mostly extract hand-crafted, application-specific location representations that heavily depend on large-scale supplementary data and thus cannot perform efficiently on mobile devices. In this paper, we present a machine learning based approach, termed GPS2Vec+, which learns rich location representations by capitalizing on the world-wide geotagged images. Once trained, the model has no dependence on the auxiliary data anymore so it encodes geotags highly efficiently by inference. We extract visual and semantic knowledge from image content and user-generated tags, and transfer the information into locations by using geotagged images as a bridge. To adapt to different application domains, we further present an attention-based fusion framework that estimates the importance of the learnt location representations under different contexts for effective feature fusion. Our location representations yield significant performance improvements over the state-of-the-art geotag encoding methods on image classification and venue annotation.

CCS CONCEPTS

• **Information systems** → **Social networks**; *Document representation*; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Location representations, pre-trained neural networks, attention-based fusion, geo-aware applications

Ying Zhang and Zhenguang Liu are the corresponding authors.
This work was done while Sheng Wang was at the National University of Singapore.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3475268

ACM Reference Format:
Yifang Yin, Ying Zhang, Zhenguang Liu, Yuxuan Liang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2021. Learning Multi-context Aware Location Representations from Large-scale Geotagged Images. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475268

1 INTRODUCTION

The geotags that are associated with multimedia documents have long been considered as an important context in a variety of applications including image classification [21], land use recognition [5], venue semantic annotation [14, 26], *etc.* A typical use scenario of geotags on mobile devices is to select candidates based on their distance to the user location. More advanced geotag encoding techniques have been proposed, but all have drawbacks that hinder their effectiveness in geo-applications. As summarized in Table 1, the existing geotag encoding techniques can be roughly classified into the following four categories. Firstly, the OneHot encoding technique [5] is widely adopted due to its simplicity, which typically segments the 2D space into cells to generate a vector indicating which cell the GPS coordinate falls into. However, it does not encode any semantics on locations so the performance gain introduced by OneHot encoding is usually quite limited.

Next, towards encoding semantics on GPS coordinates, researchers have leveraged a variety of supplementary data sources to extract semantic contexts. GIS-based geotag encoding techniques [12, 20] utilize geographical information system databases, *e.g.*, GeoNames [9] and Google Maps [10], to extract geographic features by retrieving nearby place entities. By adopting spatial indexing such as KD-tree and ball-tree [8] for efficient nearest neighbor search in the 2D space, the GIS-based encoding can be greatly speeded up. However, the supplementary GIS databases are required during the encoding, leading to additional maintenance cost. Moreover, the data completeness of GIS databases usually varies significantly in different areas of the world, making the GIS-based geotag encoding techniques only applicable to small areas of interest due to the limitations of data availability.

Thirdly, image-based geotag encoding techniques replace the GIS database by a large-scale geotagged user generated image dataset [16, 17, 21]. Rich information such as visual and semantic contexts can be extracted from the image content and the associated

Table 1: Comparison with the previous work.

Method	Location Context	Visual Context	Semantic Context	Application Independent	AOI Independent	Realtime Extraction	Sup. Data Free
OneHot [5]	✓	✗	✗	✓	✓	✓	✓
GIS-based [12] [20]	✓	✗	✗	✗	✗	✓	✗
Image-based [16] [21]	✗	✓	✓	✗	✓	✗	✗
GPS2Vec [27]	✗	✗	✓	✗	✓	✓	✓
GPS2Vec+, **Proposed**	✓	✓	✓	✓	✓	✓	✓

user tags. However, one major drawback of image-based geotag encoding lies in the high maintenance cost of a world-wide large-scale image dataset. As the efficiency of spatial indexing drops significantly when searching for nearest neighbors in the high-dimensional visual space, image-based geotag encoding is time-consuming and usually cannot be conducted in realtime.

Finally, a machine learning based geotag encoding technique called GPS2Vec [27] has been proposed, which aims to generate GPS embeddings in realtime based on pre-trained models. This method encodes the geotags based on the user generated tags only. So it remains unclear whether the model is effective on capturing multi-modal contexts and how much is the performance gain when applying the model to applications of different domains.

In this work, we present a geotag encoding approach that has an edge in simultaneously fulfilling the following two properties: *context-rich* to adapt effectively to applications from different domains and *lightweight* to deploy efficiently on mobile devices. To achieve the goals, we propose to train neural networks to learn rich contexts by transferring knowledge from image visual content and semantic user tags into locations based on a large-scale geotagged image dataset. Though the networks are trained with visual and sematic supervision, the models have no dependence on the supplementary geotagged image dataset during inference. In our experiments, we show that the context-aware location representations learnt by our network obtain significant performance gain on two typical location-based applications, *i.e.*, image classification and venue annotation. To adapt to different application domains, we further present an attention-based multi-context fusion framework to estimate the importance of contexts in different applications. Here we summarize the key contributions of this paper as follows:

- We present a machine learning based multi-context GPS encoding technique, which simultaneously learns location, visual, and semantic contexts from large-scale user-generated geotagged images.
- We introduce attentions to perform effective multi-context fusion, as the importance of contexts may vary significantly in applications from different domains.
- The trained networks have no dependence on the supplementary geotagged images during inference, leading to realtime GPS encoding and low maintenance cost, which can be readily deployed on mobile devices.
- Extensive experiments are conducted on image classification and venue annotation tasks. Empirical results show that our proposed method outperforms the state-of-the-art geotag encoding approaches by 5% ∼ 11% in terms of the mAP.

2 RELATED WORK

The GPS coordinates associated with multimedia documents can provide rich contextual information that is crucial for, *e.g.*, information understanding, document retrieval, and personalized recommendation [21, 25, 28]. For example, fusing visual features with geotags has been shown to obtain much higher accuracy in image classification [16]. However, the direct utilization of geotags makes it difficult to be integrated with existing high-dimensional textual and visual features as GPS coordinates are fine-grained location indicators that only contains two values, namely latitude and longitude [21, 27]. As shown in Table 1, OneHot encoding is one of the most popular and straightforward method that converts a geotag into a vector representation. For example, Ye *et al.* proposed to segment an area of interest (AOI) into M cells and represent each geotag as a M-dimensional one-hot vector [25]. The non-zero entry in the one-hot vector denotes the index for the corresponding grid cell that the geotag falls into. Christie *et al.* proposed to use the UTM Zones as the grid cells to encode geotags anywhere in the world [5]. However, the number of cells to encode geotags is always limited by the computational cost and memory, resulting in significant information loss when dealing with geotags that are widespread around the world.

With the availability of geotagged auxiliary databases, GIS-based and image-based geotag encoding techniques have been proposed to describe a location based on its geographical and visual neighbors. One of the early work was proposed Joshi and Luo [12]. It encodes a location by retrieving nearby place entities from GeoNames, which is a freely available GIS database [9]. Liao *et al.* proposed to encode a geotag by tag propagation from both the geographical and visual neighbors of a given geotagged image. Tang *et al.* further proposed to use multiple supplementary databases including Google Maps, American Community Survey [1], and geotagged images to extract multi-context features to describe a location [21]. However, as the corresponding image is required when extracting the visual context features of a geotag, image-based encoding techniques have two major drawbacks: 1) the methods are tailored for image classification only, and 2) the geotag encoding cannot be performed in realtime due to the time-consuming KNN search in the high-dimensional visual space. To solve the above issues, Yin *et al.* recently proposed a machine learning based geotag encoding method, which is able to extract semantic context features for a geotag in realtime without the facilitation of any supplementary databases during the feature extraction [27]. However, as only user tags are utilized, it is still difficult for this method to outperform the state-of-the-art image classification approaches.

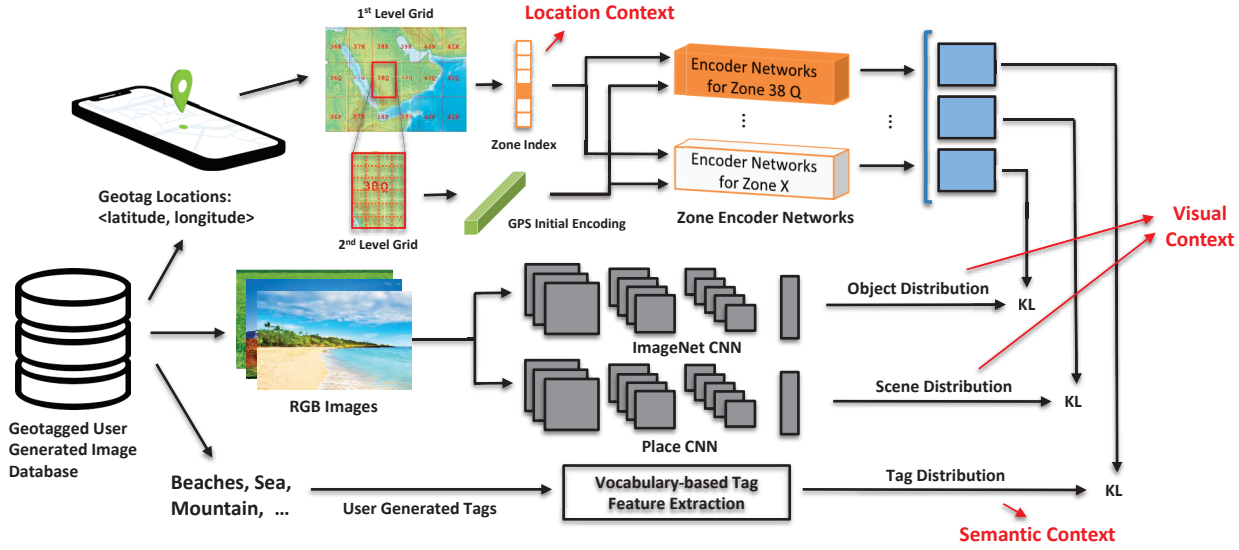


Figure 1: Overview of our proposed multi-context GPS encoding networks.

3 LEARNING MULTI-CONTEXT AWARE LOCATION REPRESENTATIONS

We propose to train *lightweight* neural networks to learn *context-rich* representations for locations worldwide by leveraging the massive amounts of user-generated geotagged images that are available online. In practice, we adopt the YLI-GEO dataset, which comprises around 6 million geotagged Flickr photos labeled with user-generated tags as well [3]. This dataset has been used in the MediaEval Benchmarking Initiative's Placing Task in 2014, 2015, and 2016 [4]. Figure 1 illustrates the system overview of our proposed geotag encoding framework. A geotagged image database is composed of three correlated components: the geotag, the RGB image, and the user tags. As it shows, we extract the location context from the geotags directly. The visual context and the semantic context of a geotag are learnt by matching the location to the object/scene distribution extracted from the image content and the tag distribution extracted from the user tags, respectively. The technical details are introduced below.

3.1 Location Context

To deal with locations worldwide, we follow the location context modeling based on UTM zones [5]. UTM refers to the Universal Transverse Mercator coordinate system, which divides the Earth into 60 longitude zones and 20 latitude bands. Each UTM zone is referenced by a longitudinal zone number (*i.e.*, 1 to 60) and a latitudinal zone letter (*i.e.*, C to X, omitting O). We generate a one-hot encoding to indicate which UTM zone the geotag falls into. This UTM-based one-hot encoding of a geotag has two usages in our framework. On one hand, it is directly considered as the location context feature of the geotag. On the other hand, we follow the two-level grid based framework to learn the visual and semantic context features in each UTM zone [27], and thus use the one-hot encoding as an index to select the correct encoder to estimate the object, scene, and tag distributions around the geotag location. Next, we introduce how we model and extract the visual and semantic contexts given a geotag in each UTM zone.

3.2 Visual Context

In each UTM zone, we train a neural network as the visual context encoder, which transfers knowledge from image content to location [2]. Let l_i be a geotag, $x_i \in \mathbb{R}^D$ denote the GPS initial encoding of l_i [27], and I_i denote the corresponding image labeled with l_i. Our goal is to use the posterior probabilities of a teacher vision network $g_k^v(I_i)$ to train our student location network $f_k^v(x_i)$ that models the distribution of the visual concepts around the input location l_i [11]. We adopt two pre-trained VGG16 teacher networks [19] trained using the ImageNet and the Places datasets [7, 29, 30] to extract the object and the scene visual concepts from images, respectively. Our student network consists of three shared hidden layers (*i.e.*, 512, 1024 and 2048) followed by ReLU (rectified linear unit) activation, and two separate output layers with softmax function to match the posterior probabilities of the two teacher networks. Let θ_v be the model parameters to be learnt, we optimize the loss function given as,

$$L(\theta_v) = \sum_{k=1}^{K} \sum_{i=1}^{N} D_{KL}(g_k^v(I_i) || f_k^v(x_i; \theta_v)) \quad (1)$$

where $D_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$ is the Kullback Leibler (KL) divergence, which measures how one probability distribution is different from a second, reference probability distribution. N is the number of geotagged images in the training dataset and K is the number of teacher vision networks we use for supervision (*i.e.*, $K = 2$). During training, we optimize the loss function using stochastic mini-batch gradient descent based on back-propagation with momentum. The mini-batch size and the momentum were set to 32 and 0.9, respectively. The learning rate was set to 0.001.

3.3 Semantic Context

Similar to the visual context modeling, we train a neural network in each UTM zone as the semantic context encoder to transfer knowledge from user tags to location. Let l_i be a geotag, $x_i \in \mathbb{R}^D$ denote the GPS initial encoding of l_i [27], and T_i denote the corresponding user tags associated with the same image. We extract a vocabulary-based tag distribution from the user tags, denoted as $g^s(T_i)$, to train our student network $f^s(x_i)$ that models the distribution of the semantic concepts around the input location l_i. Let θ_s be the model parameters to be learnt, we optimize the loss function given as,

$$L(\theta_s) = \sum_{i=1}^{N} D_{KL}(g^s(T_i)||f^s(x_i;\theta_s)) \tag{2}$$

We adopt the same student network architecture and KL divergence as the loss function, which is optimized using stochastic mini-batch gradient descent based on back-propagation with momentum during training. Next, we discuss how we extract $g^s(T_i)$, the vocabulary-based tag distribution from the user tags.

Let $V = \{t_1, t_2, ..., t_M\}$ be a vocabulary of M tags. A naive solution is to generate $\hat{g}^s(T_i)$ as an M-dimensional histogram by setting the entries of the user tags that appear in T_i to 1. However, $\hat{g}^s(T_i)$ can be sensitive to both GPS noise and user tag uncertainty. So we refine $\hat{g}^s(T_i)$ by feature propagation from the geo neighbors of l_i [16] as,

$$g^s(T_i) = \sum_{l_j \in NN(l_i)} w_{ij} \cdot \hat{g}^s(T_j) \tag{3}$$

where $NN(l_i)$ denote the geo neighbors of l_i and w_{ij} is the weight of $\hat{g}^s(T_j)$ when computing $g^s(T_i)$ by feature propagation. Follow the work proposed by Liao *et al.* [16], we compute the weight of geo neighbors based on the geographical distance between locations l_i and l_j as $\exp(-\frac{\|l_i - l_j\|_2}{\sigma})$ where σ is a constant attenuation coefficient. Finally, we normalize $g^s(T_i)$ using the L1 norm to obtain our vocabulary-based tag distribution as the semantic context feature for supervision.

3.4 Pre-trained Model based Encoding

Once the neural networks are trained, the large-scale supplementary data is no longer required. This is because that the models capture the visual and the semantic information of the supplementary data, and thus can be used as an alternative to generate the multi-context geotag encodings. In terms of the computational cost, our method divides the geotag encoding into the offline training phase (training) and the online extraction phase (inference). By doing this, we move the time-consuming processes such as the nearest neighbor search to the offline training stage. The inference in our method is performed simply by passing an input location to a neural network to output the multi-context encodings, which can be executed highly efficiently in milliseconds. In terms of the storage cost, our model is lightweight that can be easily deployed in mobile devices. Comparatively, the maintenance, backup, and recovery cost of the large-scale supplementary data is prohibitive for image-based geotag encoding approaches. Auxiliary images are mostly saved in individual files on the server side, resulting in further delays caused by the communication between the mobile and the server. Our models, on the other hand, allows true realtime response even without Internet connections.

One challenge of using pre-trained models for geotag encoding is how to process arbitrary locations anywhere in the world. In practice, it is infeasible to train one model to cover the entire Earth surface [27]. This is because that the actual GPS coordinates are very fine location indicators, which are difficult for neural networks to effectively learn from [21]. Previous work suggests to convert geotags into a dense representation based on soft one-hot encoding to be used as the input of neural networks [27]. On one hand, the information loss during this conversion is controlled by the granularity of the grid that is used for the soft one-hot encoding[1]. On the other hand, the number of cells in the grid, which equals to the dimension of the geotag dense representation, is always limited by the system's computational resources when being processed by neural networks[2]. Therefore, the use of a single grid to cover the entire planet will result in great information loss due to the coarse grid granularity. To address this issue, we follow the location context modeling based on UTM zones [5] and train one model in each UTM zone to generate the multi-context geotag encodings. Though multiple models are trained, fortunately it should be sufficient to cache one model at a time, *i.e.*, the one that covers the user location, on the mobile side in most location-aware applications. Thus, the models can be efficiently retrieved and updated locally in each UTM zone without much additional cost.

4 ATTENTION-BASED MULTI-CONTEXT FUSION

Based on transfer learning, our proposed multi-context aware location representations can introduce significant performance gain in a variety of location-aware applications. For example, in geotagged image classification, we can use our pre-trained location encoders to extract location representations from a geotag, use pre-trained ImageNet models to extract deep visual features from the corresponding image, and fuse them by concatenation to fine-tune for the target application.

When being applied in different applications, the importance of the learnt multi-context location representations can vary significantly. We thus present an attention-based feature fusion network in Figure 2, which estimates the importance of each feature and fuse the features based on their important scores rather than simply concatenating them together. Basically, we consider the following five types of features for fusion,

- One-hot location encoding (location context of a geotag)
- Object distribution histogram (visual context of a geotag)
- Scene distribution histogram (visual context of a geotag)
- Tag distribution histogram (semantic context of a geotag)
- Application-specific feature vector (*e.g.*, visual feature in image classification)

We further generate a one-hot indicator to indicate the feature type, *i.e.*, the indicator for the k-th feature is a one-hot vector with all 0 entries except the k-th entry set to 1. We estimate the feature importance based on the indicator by passing it to a fully-connected layer and the Sigmoid activation to ensure the important scores α_k to be in the range of $[0, 1]$. Features are processed by separate

[1] Geotags that fall into the same cell will be assigned with the same dense representation.

[2] A typical input size for 2D CNNs is only 224×224 for image processing.

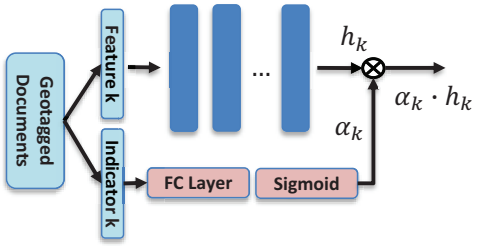


Figure 2: Attention-based multi-context feature fusion network.

sub-networks and the output h_k is fused by weighted concatenation.

$$h = \alpha_1 \cdot h_1 + \alpha_2 \cdot h_2 + \alpha_4 \cdot h_3 + \alpha_4 \cdot h_4 + \alpha_5 \cdot h_5 \quad (4)$$

where + represents the concatenation of two vectors. This strategy has been shown to be more effective than feature concatenation with equal weights as it automatically adapts to the application domain. Subsequently, we model the supervised overall loss for applications as,

$$L_{App} = L(y, \phi(\mathbf{W} \cdot h + \mathbf{b})) + \sum_{k=1}^{5} \lambda_k \cdot L(\mathbf{y}, \phi(\mathbf{W_k} \cdot h_k + \mathbf{b_k})) \quad (5)$$

where y is the ground-truth labels, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{W_k}$, and $\mathbf{b}_k$ are the trainable parameters of the output layers, ϕ and L are the application-specific activation function and loss function, respectively. For example, for a multi-label classification problem, ϕ and L denote the Sigmoid activation and the binary cross-entropy error, respectively. The second term $\sum_{k=1}^{5} \lambda_k \cdot L(\mathbf{y}, \phi(\mathbf{W_k} \cdot h_k + \mathbf{b_k}))$ is to improve the descriptiveness of h_k, which we find is beneficial to obtain potential performance gain. In our experiments, we empirically set the balancing factors λ_k to 1 and compute the final predictions, denoted as p, using the fused feature h only, *i.e.*, $p = \phi(\mathbf{W} \cdot h + \mathbf{b})$.

5 EXPERIMENTS

We evaluate the proposed method on two global-scale datasets obtained from real-world applications in different domains: geotagged Flickr images for image classification and geotagged Foursquare check-ins for venue annotation. For evaluation, we first present the experimental setups, followed by model justification and comparison to the state-of-the-arts. Finally, we perform an ablation analysis to discuss the differences and advantages of our proposed method over the existing geotag encoding approaches.

5.1 Experimental Setup

Datasets. For image classification, we evaluate our method using the NUS-WIDE dataset [6], which is a benchmark dataset widely used in image classification. We use the geo-tagged images in NUS-WIDE and form a training set with 41,173 images and a test set with 27,401 images. Ignoring rare concepts, we test on 75 concepts covering objects, scenes, and events. In terms of the visual feature, we adopt the BovW representation based on SIFT descriptors that is used in previous work [16] to make it a fair comparison. For venue annotation, we conduct experiments on a global-scale check-in dataset collected from Foursquare [23, 24]. Each check-in records user id, venue id, visit time, latitude, longitude, and venue category name. We test on six popular categories as shown in Table 2, and formulate it as a multi-label classification problem. To generate the multi-context aware location representations, we learn the visual context from the YLI-GEO dataset [3] and learn the semantic context from the 1M Flickr dataset [15] to make it a fair comparison with the previous work.

Table 2: Venue categories and their percentages (z%) in the evaluation dataset.

Category	z%	Category	z%
Restaurant & Food	40.4%	Shop	19.8%
Hotel	3.5%	Pub & Bar	13.1%
Store	20.7%	Hospital	2.5%

Parameter Settings. To extract semantic context from user tags, we adopt the same tag vocabulary V used in previous work [16, 27] to generate the tag distribution features. The number of geo neighbors is set to 150, and the attenuation coefficient σ in Eq. 3 is set to 10 [16]. In the attention-based multi-context fusion network, we process the features by separate sub-networks that consist of one dense layer with 512 hidden units followed by ReLU activation[3]. The fused representation is passed to the output layer with Sigmoid to obtain the final predictions.

5.2 Model Justification

We perform a step-by-step model justification to demonstrate the effectiveness of our proposed 1) location representations and 2) attention-based multi-context fusion framework.

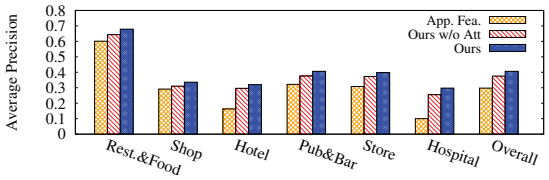
Table 3 reports the mean average precision obtained based on location representations learnt from different contexts and their fusion. The application-specific feature (*i.e.*, App. Fea.) refers to image visual feature in image classification and check-in data statistics in venue annotation, the details of which will be introduced in the next section. The location representation generated based on UTM contains no semantic information and thus performs the least effectively in both image classification and venue annotation. Comparatively, the location representations generated based on the visual context (*i.e.*, object and scene) and the semantic context (*i.e.*, tag) perform significantly better. As different contexts carry complementary information about different aspects of object, events, and activities around a location, we are able to obtain a mAP gain of 13.5% and 6.8% in image classification and venue annotation by combining them together compared to the location representation generated based on UTM only. Next, we further combine our proposed multi-context location representations with application-specific features, and we are able to obtain a mAP gain of 25.9% and 16.6% based on simple concatenation and 28.2% and 19.6% based on our proposed attention-based fusion.

Finally, we compare the per-class average precision obtained based on the application-specific feature, the fusion of the application-specific feature and our proposed multi-context geotag representations with or without the attention-based mechanism in

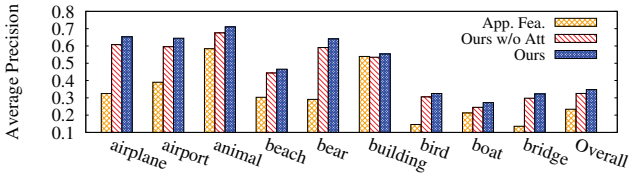
[3]Except the UTM-based location context feature is processed by a dense layer with 16 hidden units due to the limited information it contains.

Table 3: Mean average precision comparison based on geotag representations learnt from different contexts and their fusion.

Context	Att.	Image Classification		Venue Annotation	
		mAP	Gain	mAP	Gain
UTM	-	0.066	-	0.210	-
Object	-	0.191	12.5%	0.265	5.5%
Scene	-	0.194	12.8%	0.253	4.3%
Tag	-	0.182	11.6%	0.243	3.3%
Obj.+Sce.	✗	0.194	12.8%	0.272	6.2%
Obj.+Sce.+Tag	✗	0.198	13.2%	0.277	6.7%
UTM+Obj.+Sce.+Tag	✗	0.201	13.5%	0.278	6.8%
Contexts + App. Fea.	✗	0.325	25.9%	0.376	16.6%
Contexts + App. Fea.	✓	0.348	28.2%	0.406	19.6%



(a) Venue annotation



(b) Image classification

Figure 3: Per-class average precision comparison of our proposed methods to baselines.

Figure 3. For image classification, we report the results on 9 out of the 75 classes, namely airplane, airport, animal, beach, bear, building, bird, boat, and bridge. By training with application-specific feature only, we obtained a mAP of 0.234 and 0.298 in image classification and venue annotation, respectively. By further fusing with our proposed multi-context geotag representations, mAP gain can be obtained in most of the classes. Overall, our proposed method improved the mAP by 9.1% and 7.9% with concatenation-based fusion, and by 11.4% and 10.8% with our proposed attention-based fusion in image classification and venue annotation, respectively. The experimental results indicate the effectiveness of our proposed multi-context location representations and attention-based feature fusion approach.

5.3 Comparison with the State-of-the-art

In this section, we compare our proposed method against the state-of-the-art geotagged image classification and venue annotation approaches. OneHot [5] generates a one-hot encoding for a location based on grids defined by the UTM zones. HashTag [21] defines a GPS location based on a set of radii R. For each $r \in R$, it pools over a circle of radius r around that location and counts the number of images tagged with tag t that fall into the radius. TagFeature [16] defines a GPS location based on its nearest K geographical neighbors. It counts the number of images tagged with tag t in its neighborhood weighted by the geographical distance between them. Finally, GPS2Vec [27] encodes a location based on pre-trained models capturing the semantic context only. The results are reported in Tables 4 and 5 with the **best fusion-based result** and the best geo-based result highlighted.

Table 4: Mean average precision comparison on geotagged image classification.

Method	Classifier	mAP
Visual	visual	0.234
OneHot [5]	geo	0.066
HashTag [21]	geo	0.163
TagFeature [16]	geo	0.161
GPS2Vec$_{onehot}$ [27]	geo	0.130
GPS2Vec [27]	geo	0.182
Ours	geo	0.210
OneHot [5] + V	fusion	0.238
HashTag [21] + V	fusion	0.261
TagFeature [16] + V	fusion	0.260
GPS2Vec$_{onehot}$ [27] + V	fusion	0.277
GPS2Vec [27] + V	fusion	0.300
Kleban *et al.* [13]	fusion	0.080
Qian *et al.* [18]	fusion	0.113
Li *et al.* [15]	fusion	0.251
Wang *et al.* [22]	fusion	0.236
Liao *et al.* [16]	fusion	0.347
Ours + V	fusion	**0.348**

The results on image classification are reported in Table 4 where the application-specific feature refers to the image visual feature. Visual-, geo-, and fusion-based classifiers indicate the information we used for image classification. As can be seen, our proposed method obtained the best mAP of 0.210 among the geo-based methods. By fusing with the visual features, our method outperformed the state-of-the-art geotag encoding methods HashTag+V, TagFeature+V, GPS2Vec+V by 8.7%, 8.8%, and 4.8%, respectively. More

Table 5: Performance comparison on venue semantic annotation.

Method	Hamming Loss	Coverage Error	mAP
Baseline [26]	0.159	2.129	0.298
OneHot [5]	0.165	2.235	0.210
HashTag [21]	0.164	2.200	0.249
TagFeature [16]	0.165	2.198	0.247
GPS2Vec$_{onehot}$ [27]	0.165	2.224	0.220
GPS2Vec [27]	0.164	2.189	0.243
Ours	0.163	2.154	0.283
OneHot [5] + B	0.152	2.045	0.337
HashTag [21] + B	0.153	2.060	0.335
TagFeature [16] + B	0.154	2.075	0.323
GPS2Vec$_{onehot}$ [27] + B	0.151	2.034	0.340
GPS2Vec [27] + B	0.151	2.029	0.344
Ours + B	**0.148**	**1.962**	**0.406**

importantly, our method is also able to obtain better (at least competitive results) compared to the state-of-the art location-aware image classification methods. Take the method proposed by Liao *et al.* as an example, here we report the best mAP they obtained by propagating tags from both the geo neighbors and the visual neighbors of the test image[4]. This method is tailored for image classification and can hardly be realtime as it searches for the visual neighbors in the high-dimensional feature space. Comparatively, our method provides a realtime and general solution for all location-aware applications. At the same time, our method achieved a competitive mAP of 0.348, which indicates the effectiveness of our proposed approach.

The results on venue annotation are reported in Table 5 where the application-specific feature refers to the statistics extracted from the check-in records [26]. For example, users are likely to behave differently at different venues due to the nature of functions offered by these places. Therefore, different behavior patterns of visitors can be extracted from the check-in data at each venue to depict the place. According to the previous work [26], we extract a 34-D baseline feature vector based on 1) total number of visits, 2) total number of unique visitors, 3) maximum number of check-ins by a single visitor, 4) distribution of visit time in a week, and 5) distribution of visit time in a day. For performance comparison, we follow previous work [26] and report the following three metrics: hamming loss, coverage error, and average precision. Hamming loss measures the fraction of labels that are incorrectly predicted, while the rest two metrics measure the ranking performance of the predictors. As can be seen, our proposed method obtained the best mAP of 0.283 and 0.406 among the geo-based and fusion-based methods, respectively. By fusing the location representations with the baseline feature, our method outperformed the second best GPS2Vec+B by 2.0%, 3.3%, and 18.0% in terms of hamming loss, coverage error, and mean average precision, respectively.

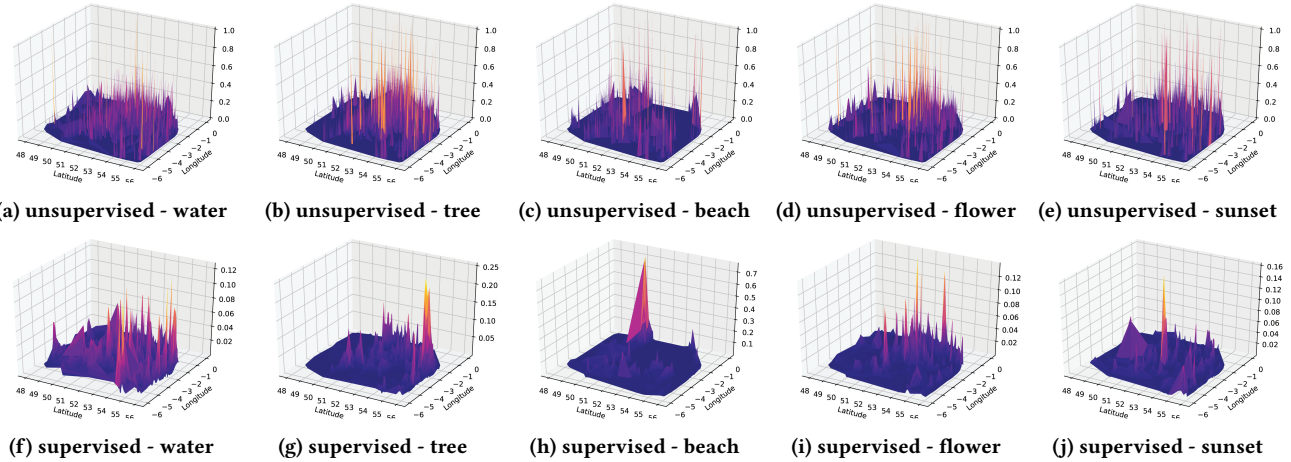
[4]TagFeature propagates from the geo neighbors only.

5.4 Visualization and Discussion

Our proposed method provides a general geotag encoding solution with the advantages of being real-time, context-rich, application-independent, AOI-independent, *etc.* In this section, we would like to specifically analyse the robustness of our proposed method to the GPS noise in the geotags and the uncertainty in the user-generated content, compared to the existing image-based geotag encoding approaches (see Table 1). Here we refer to the image-based geotag encoding as unsupervised method and refer to our proposed approach as supervised method. Take the semantic context modeling in Section 3 as an example, the unsupervised method computes a vocabulary-based tag distribution histogram as the encoding of a location by propagating information from its geo neighbors [16]. Our proposed method, on the other hand, use the vocabulary-based tag features generated by the unsupervised method as labels to train neural networks to capture the statistics of the supplementary dataset. At a first thought, one may worry that the supervised method, though can encode geotags in realtime, may perform less accurately than the unsupervised method as the neural network output cannot precisely match to the labels due to the information loss during training. However, it is worth emphasizing that our goal is to generate robust and descriptive geotag encodings rather than accurately re-generating the unsupervised method's results based on pre-trained models. In fact, though the unsupervised method aggregates the information of the geo neighbors with weights computed based on distance, both the weights and the generated tag features are hand-crafted and thus can be sensitive to data noises.

To illustrate, we visualize the vocabulary-based tag features generated by both the unsupervised method and our neural network based supervised method in Figure 4. We visualize the distribution of five tags (*i.e.*, water, tree, beach, flower, and sunset) in the UTM zone 30U. As can be seen, the tag features generated by the unsupervised method is quite noisy. Comparatively, our supervised method is able to generate more smoothed tag distributions that are more in line with the real situation. For example, the distribution of tag "beach" shows strong correlations to its true location, while the distribution of tag "water" or "flower" tends to be more evenly spread out in the geospace. Therefore, the tag features learnt by our supervised method tend to be more robust, which can achieve significant performance gain especially when fusing with the application-specific features. For verification, we select the top nine most popular UTM zones[5] and report the results obtained by fusion (*i.e.*, Tag + App. Fea.) on image classification in Table 6. As can be seen, the supervised method outperforms the unsupervised method in seven out of the nine UTM zones by a large margin. Overall, a mAP gain of 4% and 2% can be obtained by replacing the hand-crafted tag features by our learnt location representations in image classification (when fusing with image visual features) and venue annotation (when fusing with the baseline features), respectively. Thus, our proposed method not only has the advantages of being realtime without the facilitation of any supplementary datasets, but also generates more robust and descriptive location representations, compared to the existing unsupervised location encoding methods.

[5]The popularity of the UTM zones is measured based on the number of testing images located in them.



(a) unsupervised - water (b) unsupervised - tree (c) unsupervised - beach (d) unsupervised - flower (e) unsupervised - sunset

(f) supervised - water (g) supervised - tree (h) supervised - beach (i) supervised - flower (j) supervised - sunset

Figure 4: Visualization of the vocabulary-based tag features generated by both the unsupervised method and our proposed supervised method in UTM zone 30U (latitude: $48^\circ \sim 56^\circ$, longitude: $-6^\circ \sim 0^\circ$).

Table 6: Mean average precision comparison per UTM zone of the vocabulary-based tag features generated by unsupervised and supervised methods on image classification.

UTM Zones	30U	31U	18T	10S	32T	32U	11S	17T	16T
unsupervised	0.246	0.257	0.254	0.351	0.279	0.239	0.301	0.239	0.295
supervised	0.293	0.282	0.278	0.351	0.322	0.234	0.323	0.253	0.346
mAP Gain	4.7%	2.5%	2.4%	0%	4.3%	-0.5%	2.2%	1.4%	5.1%

Table 7: Efficiency (in seconds) comparison of unsupervised and supervised geotag encoding methods.

Method	Encoding time (s)
unsupervised - Geo KNN	0.008
unsupervised - Visual KNN	2.050
supervised - Ours	0.029

Finally, we analyse the efficiency of our proposed geotag encoding method. The experiments were conducted on an Intel® Xeon® E5-2630 v4 2.20GHz CPU based on Python implementation. The average encoding time (in seconds) per geotag is reported in Table 7. As Liao *et al.* [16] suggested, better location representations can be generated by propagating user tags from both the geo neighbors and the visual neighbors of an geotagged image. To this end, we utilized the "scikit-learn" library and adopted ball-tree [8] to perform efficient K Nearest Neighbor (KNN) search (K=150). As can be seen from Table 7, the efficiency of spatial indexing significantly declines when searching for visual neighbors in the high-dimensional feature space. On the other hand, the inference time (*i.e.*, geotag encoding time) of our proposed supervised method constantly equals to 0.029 s, regardless of retrieving only geo neighbors, visual neighbors, or both. This is because our method only uses the location representations generated by the unsupervised method as training labels. Thereby, the time-consuming KNN queries are executed only once in the training phrase. The geotag encoding (*i.e.*, inference phase) can be conducted highly efficiently based on the pre-trained models in our proposed method. Furthermore, when generating location representation based on geo neighbors only, our supervised method outperformed the unsupervised method in terms of mAP (see Table 6) with only a slight tradeoff in efficiency (see Table 7).

6 CONCLUSION

We propose to train neural networks to transfer knowledge from image content and user-generated tags to locations by utilizing large-scale geotagged Flickr images. The trained neural networks are capable of producing semantically rich location representations that are ready to be used in a variety of location-based applications. To adapt to different application domains, we further present an attention-based fusion framework that estimates the importance of the learnt location representations under different contexts to perform effective feature fusion in different applications. We have evaluated our proposed method on two representative location-based applications: image classification and venue annotation. Experimental results show that our proposed method outperforms the second best location encoding approach GPS2Vec by 5% and 6% on image classification and venue annotation, respectively.

7 ACKNOWLEDGMENT

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number MOE2018-T2-1-103. Rajiv Ratn Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT Delhi.

REFERENCES

[1] American Community Survey. 2020. http://www.census.gov/acs/www/.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning Sound Representations from Unlabeled Video. In *Advances in neural information processing systems*. 892–900.

[3] Julia Bernd, Damian Borth, Carmen Carrano, Jaeyoung Choi, Benjamin Elizalde, Gerald Friedland, Luke Gottlieb, Karl Ni, Roger Pearce, Doug Poland, et al. 2015. Kickstarting the Commons: The YFCC100M and the YLI Corpora. In *Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. 1–6.

[4] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. 2015. The Placing Task at MediaEval 2015. In *MediaEval*.

[5] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional Map of the World. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A Real-world Web Image Database from National University of Singapore. In *ACM International Conference on Image and Video Retrieval*. 48:1–48:9.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *IEEE CVPR*. 248–255.

[8] Mohamad Dolatshah, Ali Hadian, and Behrouz Minaei-Bidgoli. 2015. Ball*-tree: Efficient Spatial Indexing for Constrained Nearest-Neighbor Search in Metric Spaces. *arXiv preprint arXiv:1511.00628* (2015).

[9] GeoNames. 2020. http://www.geonames.org/.

[10] Google Maps. 2020. https://maps.google.com/.

[11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *IEEE CVPR*. 2827–2836.

[12] Dhiraj Joshi and Jiebo Luo. 2008. Inferring Generic Activities and Events from Image Content and Bags of Geo-tags. In *International Conference on Content-based Image and Video Retrieval*. 37–46.

[13] Jim Kleban, Emily Moxley, Jiejun Xu, and B. S. Manjunath. 2009. Global Annotation on Georeferenced Photographs. In *ACM International Conference on Image and Video Retrieval*. 12:1–12:8.

[14] John Krumm and Dany Rouhana. 2013. Placer: Semantic Place Labels from Diary Data. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 163–172.

[15] Xirong Li, Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2012. Fusing Concept Detection and Geo Context for Visual Search. In *ACM International Conference on Multimedia Retrieval*. 4:1–4:8.

[16] S. Liao, X. Li, H. T. Shen, Y. Yang, and X. Du. 2015. Tag Features for Geo-Aware Image Classification. *IEEE Transactions on Multimedia* 17, 7 (2015), 1058–1067.

[17] Hatem Mousselly-Sergieh, Daniel Watzinger, Bastian Huber, Mario Döller, Elöd Egyed-Zsigmond, and Harald Kosch. 2014. World-wide Scale Geotagged Image Dataset for Automatic Image Annotation and Reverse Geotagging. In *ACM Multimedia Systems Conference*. 47–52.

[18] Xueming Qian, Xiaoxiao Liu, Chao Zheng, Youtian Du, and Xingsong Hou. 2013. Tagging Photos Using Users' Vocabularies. *Neurocomputing* (2013), 144–153.

[19] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[20] Vincent Spruyt. 2018. Loc2Vec: Learning Location Embeddings with Triplet-loss Networks. https://www.sentiance.com/2018/05/03/loc2vec-learning-location-embeddings-w-triplet-loss-networks/.

[21] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. 2015. Improving Image Classification With Location Context. In *IEEE International Conference on Computer Vision*. 1008–1016.

[22] G. Wang, D. Hoiem, and D. Forsyth. 2009. Building Text Features for Object Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1367–1374.

[23] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. NationTelescope: Monitoring and Visualizing Large-scale Collective Behavior in LBSNs. *Journal of Network and Computer Applications* 55 (2015), 170–180.

[24] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory Cultural Mapping based on Collective Behavior in Location based Social Networks. *ACM Transactions on Intelligent Systems and Technology* 7, 3 (2016), 30:1–30:23.

[25] Di Yao, Chao Zhang, Jianhui Huang, and Jingping Bi. 2017. SERM: A Recurrent Model for Next Location Prediction in Semantic Trajectories. In *ACM International Conference on Information and Knowledge Management*. 2411–2414.

[26] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the Semantic Annotation of Places in Location-based Social Networks. In *ACM International Conference on Knowledge Discovery and Data Mining*. 520–528.

[27] Yifang Yin, Zhenguang Liu, Ying Zhang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2019. GPS2Vec: Towards Generating Worldwide GPS Embeddings. 416–419.

[28] Yifang Yin, Beomjoo Seo, and Roger Zimmermann. 2015. Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 3 (2015), 39:1–39:21.

[29] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464.

[30] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *NIPS*. 487–495.