GPS2Vec: Towards Generating Worldwide GPS Embeddings

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

Zhenguang Liu
Zhejiang Gongshang University
Hangzhou, China
liuzhenguang2008@gmail.com

Ying Zhang
National University of Singapore
Singapore, Singapore
zhangyin@comp.nus.edu.sg

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

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

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

ABSTRACT

GPS coordinates are fine-grained location indicators that are difficult to be effectively utilized by classifiers in geo-aware applications. Previous GPS embedding methods are mostly tailored for specific problems that are taken place within areas of interest. When it comes to the scale of the entire planet, existing approaches always suffer from extensive computational cost and significant information loss. To solve these issues, we present a novel two-level grid based framework to learn semantic embeddings for geo-coordinates worldwide. The Earth's surface is first discretized by the Universal Transverse Mercator (UTM) coordinate system. Each UTM zone is next processed as a local area of interest that is further divided into fine-grained cells to perform the initial GPS encoding. We train a neural network in each UTM zone to learn the semantic embeddings from the initial GPS encoding. The training labels can be automatically derived from large-scale geotagged documents such as tweets, check-ins, and images that are available from social sharing platforms. We evaluate the effectiveness of our proposed GPS embeddings in geotagged image classification. Improved classification results have been obtained based on a simple early feature fusion technique.

CCS CONCEPTS

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

KEYWORDS

GPS, semantic embedding, neural networks

ACM Reference Format:
Yifang Yin, Zhenguang Liu, Ying Zhang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. 2019. GPS2Vec: Towards Generating Worldwide GPS Embeddings. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359067

1 INTRODUCTION

Towards encoding semantics on GPS coordinates, researchers have leveraged a variety of supplementary data sources to extract semantic contexts. For example, Joshi and Luo [6] proposed to utilize GeoNames [4], which is a publicly available geographical information system database, to retrieve nearby place entities at a specific location. Tang *et al.* [12] proposed to utilize Google Maps [5] and American Community Survey [1] to extract both geographic and statistic features for locations within the United States. Liao *et al.* [9] proposed to leverage the worldwide scale geotagged images to generate a tag histogram by counting the associated user tags at a given location. However, the drawbacks of the existing methods can be summarized as follows: 1) Due to the availability of supplementary data sources, methods can only be applied within specific areas of interest; 2) High computational cost and delays caused by the maintenance of large-scale supplementary data and frequent nearest neighbor search queries; and 3) Poor versatility as methods are mostly tailored for a specific location-aware problem.

We therefore present a novel framework to learn semantic embeddings for worldwide GPS coordinates. We first adopt the UTM coordinate system to divide the Earth's surface into 60 longitude zones and 20 latitude bands. Next, we encode GPS coordinates into grid-based features and train a neural network to learn the semantic embeddings separately in each UTM zone. Previous studies mostly use only one grid to discretize the area of interest for GPS encoding [2, 14]. However, when dealing with worldwide locations, this strategy always leads to significant information loss caused by insufficient number of cells due to system computational cost limitations (*i.e.*, computer time, memory and disk space). To solve this problem, we innovatively introduce a two-level grid based approach to balance between the information loss and computational cost. Each of the UTM zone on the first level is considered as an area of interest, which is further discretized by a grid with fine granularity to perform the initial GPS encoding. While the one hot encoding is widely used in previous work [2, 12], we present a new soft GPS encoding method that relaxes the requirement on the cell size, being able to generate descriptive encoding features using grids with fewer cells. The GPS coordinates in different UTM zones may have the same initial encodings, but the extracted semantic embeddings will be different as they will be processed by

*Work done while at National University of Singapore.

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359067

different neural networks. Moreover, the use of neural networks moves the computational intensive processing, *e.g.*, nearest neighbor search queries, to the offline training stage, in order to extract the GPS embeddings real-time in the testing stage.

2 LEARNING GPS EMBEDDINGS

We propose a general solution that encodes a GPS coordinate into a semantic descriptor based on a Neural Network, and evaluate its utilization in geotagged image classification.

2.1 GPS Initial Encoding

It is difficult to directly use the GPS coordinates as the input of a neural network. Therefore, we transform the low-dimensional GPS coordinates into high-dimensional distributed vector representations before passing them to our proposed neural network to learn the semantic embeddings. Additionally, to deal with locations worldwide, we present a novel two-level grid based GPS encoding approach. On the first level, we adopt the UTM coordinate system and divide 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). A location in a zone is represented by the projected easting and northing planar coordinate pair. Considering the granularity of the UTM zones might be too coarse, we further divide each zone into $m \times m$ grid on the second level to perform the initial GPS encoding. Formally, let $Z = \{g_{ij} | i, j = 1, 2, ..., m\}$ denote the set of cells in zone Z. Let $c_{ij} = (x_{ij}, y_{ij})$ denote the center UTM coordinate of cell g_{ij}. Then for any GPS in the same zone Z, we first represent it by the corresponding UTM coordinate $l = (x, y)$. Next, we compute its initial encoding $E^l = \{e^l_{ij} | i, j = 1, 2, ..., m\}$ as,

$$e^l_{ij} = \exp(-\frac{\left\|l - c_{ij}\right\|_2}{\sigma}) \tag{1}$$

where $\left\|l - c_{ij}\right\|$ denotes the Euclidean distance between the UTM coordinates l and c_{ij}, and σ is a constant attenuation coefficient.

Existing grid-based GPS encoding methods mostly use a single grid to construct an indicator vector that indicates which grid cell the GPS coordinate falls into, resulting in a sparse feature vector with only one entry set to one [2, 12, 14]. The grid granularity is required to be very fine in order to reduce the information loss as GPS coordinates that fall into the same cell will be assigned with the same encoding feature. When it comes to the scale of the entire planet, the use of a single grid may result in great information loss as the number of cells is always limited by the system's computational resources. Comparatively, our approach processes each UTM zone individually by introducing multiple grids in order to improve the system's scalability. Moreover, our soft encoding approach using Eq. 1 can better discriminate GPS coordinates as the distance $\left\|l - c_{ij}\right\|$ is more sensitive to the location change in the corresponding UTM coordinate l.

2.2 Vocabulary-based Semantic Feature

Given the GPS initial encoding features, we aim to learn GPS semantic embeddings based on neural networks. The labels for training can be automatically generated by extracting semantic contexts from supplementary data sources such as Flickr, Twitter, and Foursquare. Formally, let $V = \{t_1, t_2, ..., t_n\}$ denote a vocabulary consisting of n words. Our goal is to automatically generate a vocabulary-based feature for a GPS coordinate based on vocabulary V. The resulting n-dimensional feature, denoted as $S^l = \{s^l_i | i = 1, 2, ..., n\}$, will be used as labels for the training of the neural networks.

Let $l(o)$ and $T(o)$ represent the geo tag and semantic words associated with a multimedia document o, respectively. For example, $T(o)$ can be the user tags associated with an image, the texts of a tweet, or the venue type in a check-in record. For each multimedia document, we compute its semantic encoding $S(o)$ based on vocabulary V as,

$$s_i(o) = \begin{cases} 1 & t_i \in T(o) \\ 0 & t_i \notin T(o) \end{cases} \tag{2}$$

where $s_i(o)$ is the i-th element in vector $S(o)$. This semantic encoding can be quite sensitive to both GPS noise and semantic keyword uncertainty, and therefore cannot be directly used as the vocabulary-based semantic feature. For instance, images that are geographically close to each other can sometimes have completely different user tags. To reduce noise, we smooth the geographical distribution of semantic words by taking the geo neighbors into consideration as well. Given a location l, We first retrieve the k nearest geo neighbors $NN(l)$ of location l from the geo-tagged supplementary dataset in terms of the geographical distance. Next, we compute the weighted sum of the semantic encodings in the geo neighborhood,

$$\tilde{s}^l_i = \sum_{o \in NN(l)} w^l_i(o) \cdot s_i(o) \tag{3}$$

and apply l_1 normalization to obtain our vocabulary-based semantic feature S^l.

$$s^l_i = \frac{\tilde{s}^l_i}{\sum_j \tilde{s}^l_j} \tag{4}$$

Weight $w^l_i(o)$ is formulated based on the geographical distance between locations l and $l(o)$ as [9],

$$w^l_i(o) = \exp(-\frac{\|l - l(o)\|_2}{\sigma_w}) \tag{5}$$

where $\|l - l(o)\|_2$ computes the Euclidean distance between the UTM coordinates l and $l(o)$, and σ_w is constant attenuation coefficient.

The generated feature vector S^l captures the distribution of semantic words around location l, which provides rich contextual information about events that occur in the real-world. The l_1 normalization is applied to reduce the impact caused by the unbalanced geographical distribution of the geo-tagged documents.

2.3 Neural Network Architecture

For any GPS coordinate on Earth, one may argue that it is possible to generate the corresponding vocabulary-based feature based on the unsupervised method introduced in Section 2.2 without training neural networks. However, the unsupervised method performs frequent nearest neighbor search from a worldwide large scale dataset, leading to high computational cost and processing delays.

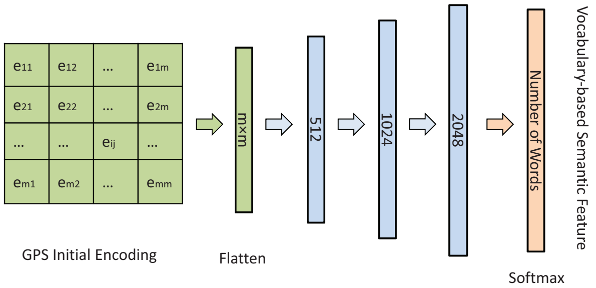

Figure 1: Illustration of the proposed neural network for GPS semantic encoding.

To solve the issues, we propose to train a neural network for each UTM zone, which is able to transform the GPS initial encoding to the vocabulary-based semantic feature. The advantages of our proposed approach are twofold: 1) the large scale supplementary dataset is only required during training, and 2) the models, once trained, generate the semantic embeddings in real-time.

As illustrated in Figure 1, we adopt a neural network that consists of three hidden layers followed by ReLU (rectified linear unit) activation, and one output layer with softmax function. The size of the three hidden layers are 512, 1024 and 2048, respectively. We leverage the one million Flickr images collected by Li *et al.* [8] to generate the vocabulary V. Following previous work [9, 13], we construct the V by selecting the top 2000 most frequent tags in the one million Flickr dataset with stop words, camera brands, and non English words excluded beforehand. So both the vocabulary-based semantic feature and the output of our neural network have a dimensionality of 2000.

The input to our neural networks is the GPS initial encoding E^l as introduced in Section 2.1. During learning, we aim to use the vocabulary-based semantic features S^l as labels to train our neural networks $f(E^l)$ to estimate the normalized word frequency in the vicinity of location l. Let θ be the model parameters to be learned, the loss function for the network training is given as,

$$L(\theta) = \sum_{l} D_{KL}(S^l || f(E^l; \theta)) \qquad (6)$$

where $D_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}$ represents the Kullback Leibler (KL) divergence. As the semantic feature S^l can be interpreted as a distribution of the semantic words in vocabulary V, the KL divergence, which measures how one probability distribution is different from a second, reference probability distribution, can be a good choice for the loss function. During training, we optimize $\arg\min_\theta L(\theta)$ 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 EXPERIMENTS

We evaluate the effectiveness of our proposed GPS semantic embeddings learning method in geo-tagged image classification. The attenuation coefficient σ is empirically set to 20 km throughout the experiments. The number of geo neighbors k is set to 150 as suggested by Liao *et al.* [9].

Table 1: mAP comparison on image classification using GPS embeddings generated based on grids with different sizes.

Grid Size	10×10	20×20	30×30
GPS2Vec	0.1754	0.1824	**0.1853**
GPS2Vec+V	0.2971	**0.2999**	0.2987

Table 2: mAP comparison with the state-of-the-art location-aware image classification approaches.

Method	Classifier	mAP
Visual	visual	0.234
OneHot [2]	geo	0.066
GPS2Vec$_{onehot}$	geo	0.130
GPS2Vec	geo	0.182
OneHot [2] + V	fusion	0.238
GPS2Vec$_{onehot}$ + V	fusion	0.277
GPS2Vec + V	fusion	0.300
Kleban *et al.* [7]	fusion	0.080
Qian *et al.* [10]	fusion	0.113
Li *et al.* [8]	fusion	0.251
Wang *et al.* [13]	fusion	0.236
Liao *et al.* [9]	fusion	**0.347**

3.1 Experimental Setup

We evaluate our method using the NUS-WIDE dataset [3]. As the location context is required in our experiments, 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. In terms of the visual feature, we adopt the BovW representation based on SIFT descriptors that is used in previous work [9] to make it a fair comparison. We adopt a neural network with one hidden layer of 512 units as the classifier. The learning rate and mini-batch size were set to 0.001 and 32, respectively. We report the mean Average Precision (mAP) as the evaluation criteria.

3.2 Performance Comparison

The image classification results obtained based on different grid sizes for the GPS initial encoding are reported in Table 1. The GPS2Vec+B method concatenated the GPS semantic embeddings and the image visual features to train a classifier. As can be seen, the GPS2Vec method obtained the best mean average precision with $m = 30$, which outperformed $m = 10$ and $m = 20$ by 5.6%, 1.6%, respectively. The GPS2Vec+B method, on the other hand, obtained competitive classification results with $m = 10$, 20, and 30. Generally speaking, the GPS semantic embeddings tend to be more descriptive with a larger grid size. Fortunately, the GPS embeddings and the visual features are complementary to each other so that the influence of the grid size has been significantly reduced with our feature fusion approach. We consider $m = 20$ can be a good trade-off, and use this setting in the next experiment.

Next, we compare our proposed method to the state-of-the-art geo-based and fusion-based image classification systems in Table 2 with the **best** and second best results highlighted. Our GPS2Vec obtained the best mAP among the geo-based methods. Moreover, the GPS2Vec+V outperformed the OneHot+V and GPS2Vec$_{onehot}$+V by 26.0% and 8.3%, respectively. The results verify the effectiveness of our proposed GPS2Vec embeddings and its complementarity to image visual features. When comparing to other fusion methods, our method achieved significant improvements over the approaches that utilized the GPS coordinates in a traditional way for geo neighbor search [7, 8, 10]. Wang *et al.* [13] and Liao *et al.* [9] both proposed to fuse a visual classifier with a textual classifier built upon tag features generated by conjunctively considering geo and visual neighbors from a supplementary image dataset. The method proposed by Liao *et al.* [9] was able to achieve the best classification result due to the following reasons. First, this method searches for visual neighbors of test images, which is tailored for image classification and cannot be applied to other geo-aware applications. Second, the authors leveraged a much larger supplementary dataset that consists of 10 million geo-tagged images to generate a more descriptive tag-based feature. Comparatively, our method is more general, and at the same time, is able to obtain the second best mean average precision of 0.3 in image classification. Moreover, our method moves the time-consuming nearest neighbor queries to the offline training stage, in order to achieve the real-time response in the testing stage of extracting semantic embeddings from GPS coordinates.

4 RELATED WORK

With the ubiquity of sensor-equipped cellphones, it is common for multimedia documents posted online to be associated with geo-tags [15]. The direct utilization of GPS coordinates makes it difficult to be integrated with existing high dimensional visual features. In recent years, several efforts have been made to encode GPS coordinates at the feature level [9, 12]. Tang *et al.* [12] proposed to grid the area of interest (AOI) into 25×25 km square cells and construct an indicator vector that indicates which grid cell the GPS coordinate falls into. Yao *et al.* [14] further proposed to transform the sparse indicator vector into a dense embedding vector by introducing an embedding layer in their system architecture. However, the number of cells to encode GPS coordinates is always limited by the computational time and memory. In an extreme example, the geotagged documents are spread all over the world [2]. The authors adopted the UTM Zone for GPS encoding, resulting in significant information loss as the granularity of UTM is too coarse.

With supplementary data sources, it is possible to encode GPS coordinates into feature vectors with semantics. For example, given a GPS coordinate, Tang *et al.* [12] extracted geographic map features and ACS features using Google Maps [5] and American Community Survey (ACS) [1], respectively. Joshi and Luo [6] proposed to encode GPS coordinates by retrieving nearby place entities from GeoNames [4], which is a freely available geographical information system (GIS) database. Recently, Vincent Spruyt [11] presented a triplet network to learn a metric space that captures semantic similarity between different geographical location coordinates. Given a location coordinate and a radius, they queried their GIS database to obtain a large amount of geographical information, and rasterized it into image tiles for the triplet network training. However, one major drawback of these approaches lies in its difficulties of being generalized to worldwide applications.

5 CONCLUSION

We have presented a novel framework, GPS2Vec, to learn GPS semantic embeddings in support of location-aware applications. The generated semantic embeddings can be easily integrated with existing high dimensional descriptors, *e.g.*, image visual features, by early fusion, based on which a new classifier can be trained to obtain more robust predictions. To divide the Earth's surface into smaller areas of manageable scale, we adopt the UTM coordinate system and train a neural network for each UTM zone to generate location semantic embeddings. Our generated GPS semantic embeddings are complementary to the textual and visual features in existing systems. The geotagged image classification demonstrates the effective use of our proposed GPS semantic embeddings in machine learning based systems.

ACKNOWLEDGMENT

This research has been supported by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number MOE2018-T2-1-103.

REFERENCES

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

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

[3] 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.

[4] GeoNames. [n.d.]. http://www.geonames.org/.

[5] Google Maps. [n.d.]. https://maps.google.com/.

[6] Dhiraj Joshi and Jiebo Luo. [n.d.]. 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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/.

[12] 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.

[13] 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.

[14] 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.

[15] 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.