Multi-modal Representation Learning for Successive POI Recommendation

Lishan Li ls-li14@mails.tsinghua.edu.cn
Ying Liu[*] liuying@cernet.edu.cn
Jianping Wu jianping@cernet.edu.cn
Lin He helin1170@gmail.com
Gang Ren rengang@cernet.edu.cn
Tsinghua University, Beijing, 100084, China

Abstract

Successive POI recommendation is a fundamental problem for location-based social networks (LBSNs). POI recommendation takes a variety of POI context information (e.g. spatial location and textual comment) and user preference into consideration. Existing POI recommendation systems mainly focus on part of the POI context and user preference with a specific modeling, which loses valuable information from other aspects. In this paper, we propose to construct a multi-modal check-in graph, a heterogeneous graph that combines five check-in aspects in a unified way. We further propose a multi-modal representation learning model based on the graph to jointly learn POI and user representations. Finally, we employ an attentional recurrent neural network based on the representations for successive POI recommendation. Experiments on a public dataset studies the effects of modeling different aspects of check-in records and demonstrates the effectiveness of the method in improving POI recommendation performance.

Keywords: Representation learning, attention mechanism, POI recommendation, location-based social network

1. Introduction

Location-based social network (LBSNs), such as Foursquare and Gowalla, are increasingly popular, which uses GPS features to locate users and let users broadcast their locations and comments from their mobile device. The collected huge amount of location data from millions of individuals make it possible to study human mobility patterns and understand users' preferences (Sun et al., 2017; Zhao et al., 2018). An intriguing use of location data is Point-of-Interest (POI) recommendation, namely recommending successive POIs based on users' check-in records, which attracts great research interest in recent years.

There are several prominent features in POI recommendation, such as POI attributes (spatial coordinates, textual comment) and the target user. For example, users' textual comments could represent POIs' characteristics (Gao et al., 2015; Liu et al., 2016b; Chang et al., 2018), the spatial distance close POIs might be visited together (Zhao et al., 2017; Wang et al., 2018) and the target user's preference also affects the successive POI recommendation (Yang et al., 2017b).

[*] Ying Liu is the corresponding author.

Existing approaches that exploit POI recommendation roughly falls into two paradigms. The first kind of method mainly leverages the POI context information, such as the sequential, spatial and textual correlations among POIs. Wang et al. (2018) studies the geographical influence based on POIs' geo-influence, geo-susceptibility and physical distance. Chang et al. (2018) proposes a content-aware POI representation learning method, which utilizes POI's textual comment and co-occurrences to capture POIs' characteristics. Sequential POI patterns are also studied to explore the sequential correlations (Chen et al., 2014; Kong and Wu, 2018). The second kind of method studied user preference (Yang et al., 2017b,a) over POIs from a user's historical check-in records. Yang et al. (2017b) learns user embeddings from historical trajectories, which further work as preference vectors. Yang et al. (2017a) proposes a semi-supervised learning framework from user and POI context graph to regularize user preferences. Although aforementioned methods achieve some success, they are very specific to model part of POI recommendation features, such as only modeling spatial POI connections without considering POIs' textual comment or only modeling POIs' context feature without considering users' general preferences. All these POI features contain valuable information for successive POI recommendation, while it's challenging to characterize multiple aspects in a unified way.

In this paper, we propose a multi-modal representation (MMR) learning approach, which captures aforementioned POI features uniformly and is easy to expand more possible characteristics. Firstly, we construct a multi-modal check-in graph (MMCG) which contains heterogeneous nodes (users/POIs) and edges. The heterogeneous edges model POI correlations from five perspectives, including co-occurrence, spatial, textual location-location edge that models POI context, user-location and user-user edge that models user preference. Secondly, inspired by recent success of word embedding (Mikolov et al., 2013) and network embedding (Tang et al., 2015; Cui et al., 2018) approaches, we jointly learn POI/user representations from multi-modal check-in graph by capturing second-order proximity. Thirdly, we further propose an attentional recurrent neural network based on user and POI/location representations to recommend successive POI. Our major contributions are summarized as follows:

- We formally define a heterogeneous multi-modal check-in graph, which combines multiple aspects from check-in records in a unified way.
- We propose multi-modal representation learning approach, which jointly learns co-occurrence, spatial, textual context, and also models user preference from MMCG.
- We further propose an attentional recurrent network based on MMR for POI recommendation.
- We conduct extensive experimental evaluations based on a public dataset. The evaluation results demonstrate the effectiveness of our proposed method.

2. Related Work

2.1. Successive POI Recommendation

Conventional POI recommendation relies on POI context information to predict successive POIs. Chen et al. (2014); Zhang et al. (2014) leverages Markov Chain model or Hidden

Markov Chain model to exploit users' check-in sequences. Some research work extends POI recommendation models by considering temporal characteristics. Zhao et al. (2016) proposes a spatial-temporal latent ranking (STELLAR) method to capture the temporal influence of times on POI recommendation. However, it focuses on POI pairs and cannot model the whole check-in sequence. ST-RNN (Liu et al., 2016a) extends the RNN model and takes temporal and spatial contextual information into consideration. The model replaces the single transition matrix in RNN with time-specific transition matrices and distance-specific transition matrices. HST-LSTM (Kong and Wu, 2018) combines spatial-temporal influence in LSTM network to predict successive POI. POIs' textual comments can also provide very useful information for POI recommendation task. Gao et al. (2015) investigates various types of content information on LBSNs in terms of sentiment indications, user interests, and POI properties, which are incorporated into a unified POI recommendation framework. However, these models randomly initialize vectors as POI representations. Thus, they cannot efficiently utilize POIs' specific characteristics and relationships between POIs.

2.2. POI Embedding Learning

There have been many research efforts that utilize embedding technique to learn POI representation for POI recommendation task. Liu et al. (2016b) learns a POI's latent representation vector by leveraging the Skip-gram model. Moreover, the model is extended by considering temporal influence to train a time latent representation vector. Feng et al. (2017) proposes a new latent representation model, namely POI2Vec, to incorporate geographical influence. The geographical influence is reflected by the physical distance between POIs based on hierarchical binary tree. Zhao et al. (2017) proposes a temporal POI sequential embedding model to capture the contextual check-in information and temporal characteristics as well. This work shows that check-in sequences in different days exhibit low correlation. Chang et al. (2018) proposes a content-aware embedding model, which utilizes both check-in sequence information and text content. The injection of text content is beneficial to capture contextual characteristics of POIs. Furthermore, there have been some research work (Yang et al., 2017b,a) that incorporate user preference into POI representation by embedding techniques. In addition, embedding learning techniques have also been successfully used in other location-based social network issues, including trajectory similarity computing (Li et al., 2018) and trajectory clustering (Yao et al., 2017, 2018) and social circle inference (Gao et al., 2018). Although above methods achieve some success, they are very specific to model part of POI recommendation features. They cannot characterize multiple aspects in a unified way.

3. Preliminaries

Definition 1 ***(Check-in)*** *Let $\mathcal{U}$ denote a set of unique users, $\mathcal{L}$ denote a set of locations (a.k.a check-in points or POIs) and $\mathcal{T}$ denote the time domain. A check-in c is a triple $\langle u, l, t\rangle \in \mathcal{U} \times \mathcal{L} \times \mathcal{T}$, which means the user u has visited l at time t. Specifically, each location $l \in \mathcal{L}$ has the coordinate information of longitude and latitude (lon, lat) and users' textual comment s.*

Definition 2 ***(User-trajectory)*** *Let $\mathcal{C}$ be a collection of check-ins. Given a user u, his/her trajectory is a sequence of triplets related to u: $\mathcal{C}_u = \{\langle u, l_1, t_1 \rangle, \cdots, \langle u, l_i, t_i \rangle, \cdots \langle u, l_N, t_N \rangle\}$, where N is the sequence length and the triplets are ordered by time ascendingly.*

Definition 3 ***(Multi-modal Check-in Graph(MMCG))*** *A multi-modal check-in graph is a heterogeneous undirected weighted graph $G = (V, E)$, where $V = \mathcal{U} \cup \mathcal{L}$ is a set of heterogeneous nodes and $E = E_{clcl} \cup E_{slsl} \cup E_{tltl} \cup E_{ul} \cup E_{uu}$ is set of heterogeneous edges including co-occurrence location-location edges E_{clcl}, spatial location-location edges E_{slsl}, textual comment location-location edges E_{tltl}, user-location edges E_{ul} and user-user edges E_{uu}.*

Figure 1 gives an example of a multi-modal check-in graph. We obtain edges of G from user-trajectory and location attributes. The weight of co-occurrence location-location edge is the number of times that two locations co-occur in the user-trajectory of a given context window size. The weight of spatial location-location edge represents how close two locations are. The weight of textual comment location-location edge indicates the semantic similarity of two locations from users' perspectives. The weight of user-location edge indicates a user's general preference over each location. User-user edge represents how similar of two users' preferences over locations.

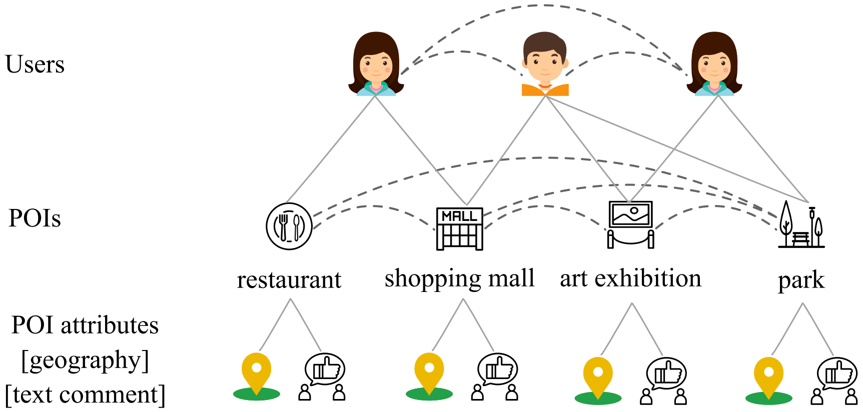


Figure 1: An example of the multi-modal check-in graph.

<u>Problem Statement:</u> Given a user $u \in \mathcal{U}$ and his/her trajectory history C_u, the goal of successive POI recommendation is to recommend a ranked list of POIs that the user u is likely to visit next at each step.

4. Multi-modal Recurrent Recommendation Framework

We study the successive POI recommendation problem via multi-modal representation learning. Figure 2 presents the overall architecture of proposed multi-modal recurrent recommendation framework. It consists of three major components: (1) multi-modal check-in graph construction, (2) multi-modal representation learning and (3) attentional recurrent network for recommendations.

4.1. Multi-modal Check-in Graph Construction

We construct a multi-modal check-in graph based on user-trajectory data and location attributes (i.e. spatial coordinates and textual comment). The nodes set V is the union set of users $\mathcal{U}$ and locations $\mathcal{L}$. The challenging issue here is how to measure the weights of the set of heterogeneous edges E. We will model each edge type as follows:

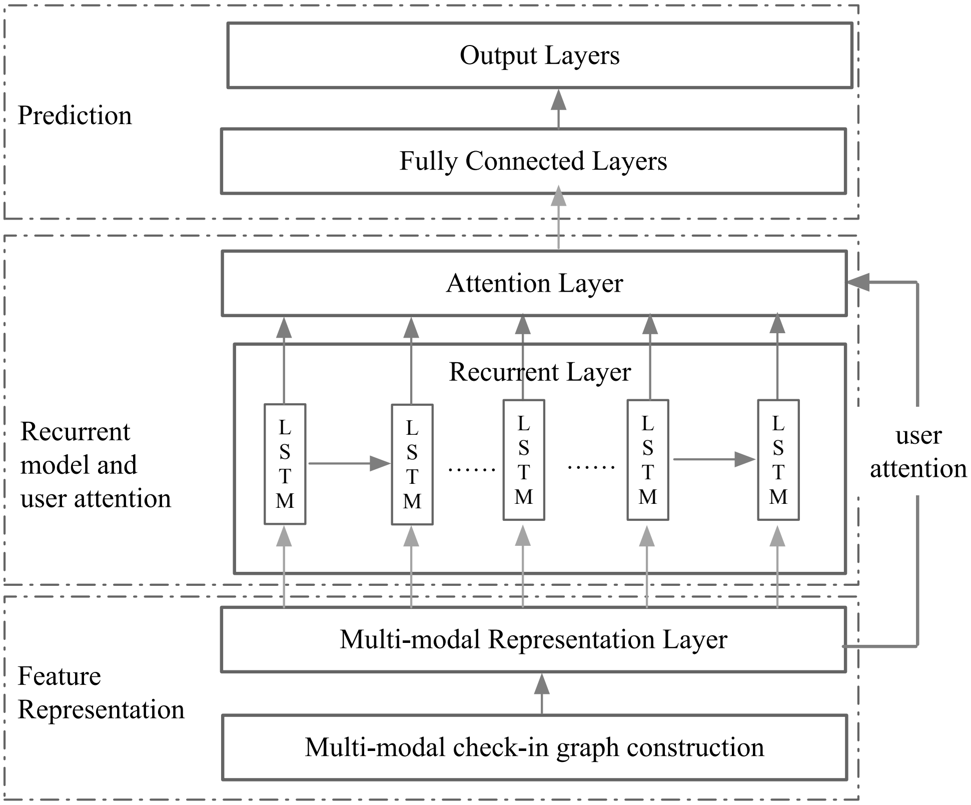


Figure 2: The overall architecture of multi-modal recurrent recommendation framework.

Co-occurrence Location-Location Edge captures the location co-occurrences in local context, which is essential information used in many embedding based approaches, such as Skip-gram (Mikolov et al., 2013). The co-occurrence explicitly reflects how frequent two locations occur in a user-trajectory. We define a data-driven approach to quantify the weight between two locations:

$$weight_{cl}(l_i, l_j) = \frac{\sum_{\mathcal{C}_u \in \mathcal{C}} F(C_u, l_i, l_j, b)}{\sum_{\mathcal{C}_u \in \mathcal{C}} \sum_{l_k \in \mathcal{L}} (F(C_u, l_i, l_k, b) + F(C_u, l_j, l_k, b))} \tag{1}$$

where $F(\cdot)$ computes the frequency of locations l_i and l_j co-occurred in user trajectory C_u with context window size b.

Spatial Location-Location Edge represents the spatial connections between two locations. Intuitively, after users visit one location, they might continue to visit its spatial close locations. We measure the edge weight based on the geographic distance as follows:

$$weight_{sl}(l_i, l_j) = \frac{1}{dist(l_i, l_j)} \tag{2}$$

$$dist(l_i, l_j) = \sqrt{(lon_i - lon_j)^2 + (lat_i - lat_j)^2} \tag{3}$$

This weight is aligned with the property that the closer two locations, the higher their associated weight.

Textual Location-Location Edge denotes the semantic relationship between two locations from users' perspectives. Users may give comments on locations after they have visited them. The textual comments reflect users' preferences on the locations. If two locations have semantic similar comments, they may share similar qualities that attracts users to visit, such as price and taste. We measure the weight as the cosine similarity of their comment representations:

$$weight_{tl}(l_i, l_j) = \frac{\sum_k d_{ik} d_{jk}}{\sqrt{\sum_k d_{ik}^2}\sqrt{\sum_k d_{jk}^2}} \tag{4}$$

where d_i and d_j are the textual comment representation vectors of locations l_i and l_j, respectively. Here we used pretrained word embeddings (Pennington et al., 2014) and obtained aggregated comment document representation.

User-Location Edge reflects users' general preferences over locations. Some users may prefer locations in recreation area, such as shopping mall. Some other users might prefer spending their time at locations around working area, such as Starbucks. The weight between user u and location l is defined based on the frequency of l in user u's trajectory $\mathcal{C}_u$.

$$weight_{ul}(u, l) = \frac{Count(\mathcal{C}_u, l)}{N} \tag{5}$$

where $Count(\cdot)$ counts the number of times user u has visited location l and N is the user trajectory length $\mathcal{C}_u$.

User-User Edge represents how similar two users' general preferences. If two users have similar preference, the POIs one user has visited might also attract another user to visit. The weight between users u_i and u_j is defined as follows:

$$weight_{uu}(u_i, u_j) = \frac{\sum_{l_k \in \mathcal{C}_{u_i} \cup \mathcal{C}_{u_j}} Count(\mathcal{C}_{u_i}, l_k) \times Count(\mathcal{C}_{u_j}, l_k)}{\sqrt{\sum_{l_k \in \mathcal{C}_{u_i}} Count(\mathcal{C}_{u_i})^2}\sqrt{\sum_{l_k \in \mathcal{C}_{u_j}} Count(\mathcal{C}_{u_j})^2}} \tag{6}$$

where $Count(\cdot)$ counts the number of times user u has visited location l.

4.2. Multi-modal Representation Learning

Multi-modal check-in graph (MMCG) contains valuable information in describing users' and locations' characteristics. Inspired by network embedding approaches (Tang et al., 2015; Cui et al., 2018), we propose multi-modal representation learning model (MMR) to train user and location embeddings from MMCG. MMR captures not only the co-occurrence, spatial and textual influences of POIs, but also the user general preference characteristics. MMR consists of location/POI context modeling and user preference modeling.

Location Context Modeling

Considering that co-occurrence, spatial and textual location-location edges characterize different aspects of locations. We use three embedding vectors $v_c, v_s, v_t$to represent the location embeddings with different edge types, respectively. And concatenate three embedding vectors as the final location embedding vector $v_l = [v_c; v_s; v_t]$.

The general idea of learning the embedding vectors is to preserve the second-order proximity from the graph (Tang et al., 2015). Given a graph $G = (V, E)$, we define the conditional probability of node n_i generated by node n_j as:

$$p(n_i|n_j) = \frac{exp(v_i^T \cdot v_j)}{\sum_{k \in V} exp(v_k^T \cdot v_j)} \tag{7}$$

where v_i and v_j are the embedding vectors of node n_i and n_j, respectively. To preserve the graph structure, we can make the conditional distribution $p(\cdot|v_j)$ be close to its empirical distribution $\hat{p}(\cdot|v_j)$, which can be achieved by minimizing their KL-divergence distance:

$$\mathcal{O} = \sum_{j \in V} d(\hat{p}(\cdot|v_j), p(\cdot|v_j)) \tag{8}$$

where $d(\cdot, \cdot)$ is the KL-divergence between two distributions. Besides, the empirical distribution of $\hat{p}(v_i|v_j)$ is set as $\frac{w_{ij}}{\sum_i w_{ij}}$. Then the objective function is minimize:

$$\mathcal{O} = -\sum_{(i,j) \in E} w_{ij} logp(v_i|v_j) \tag{9}$$

Based on above analysis, we treat Co-occurrence, spatial and textual location-location edges separately. In other words, these edges construct sub-graphs with same edge types. The objective function for each edge type is defined as follows:

$$\mathcal{O}_{cl} = -\sum_{(i,j) \in E_{clcl}} weight_{cl}(l_i, l_j) logp(v_{c_i}|v_{c_j}) \tag{10}$$

$$\mathcal{O}_{sl} = -\sum_{(i,j) \in E_{slsl}} weight_{sl}(l_i, l_j) logp(v_{s_i}|v_{s_j}) \tag{11}$$

$$\mathcal{O}_{tl} = -\sum_{(i,j) \in E_{tltl}} weight_{tl}(l_i, l_j) logp(v_{t_i}|v_{t_j}) \tag{12}$$

User Preference Modeling

User embedding representation models the general preference of users over locations. Hence user embedding could be used as a guide for recommending successive POIs for users based on their user-trajectories. In MMCG, there are two user-related edges, namely user-location edges and user-user edges. Here we employ the unique user representation for two edge types since they all characterize the users' preferences. The objective function is to minimize:

$$\mathcal{O}_{ul} = -\sum_{(i,j) \in E_{ul}} weight_{ul}(u, l) logp(v_u|v_l) \tag{13}$$

$$\mathcal{O}_{uu} = -\sum_{(i,j)\in E_{uu}} weight_{uu}(u_i, u_j) logp(v_{u_i}|v_{u_j}) \tag{14}$$

At last, the final objective function for multi-modal representation learning is minimizing:

$$\mathcal{O} = \mathcal{O}_{cl} + \mathcal{O}_{sl} + \mathcal{O}_{tl} + \beta(\mathcal{O}_{ul} + \mathcal{O}_{uu}) \tag{15}$$

where β is the hyperparameter that controls the influence of user preference modeling.

4.3. Attentional Recurrent Network

The recurrent network aims to capture the complicated sequential information or long-term dependencies contained in the user-trajectory. The recurrent layer takes the location vector sequence embedded by the multi-modal representation layer as input and outputs the hidden state step by step. The hidden states are regarded as the current status of the trajectory. Here we choose Long Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997) as the basic recurrent unit. In addition, the attention mechanism (Fan et al., 2018) is designed to make the important location hidden states contribute more on the final POI recommendation with the guidance from user embedding vector. In other words, the user's general preference affects how the locations in historical trajectory decides the successive POIs.

Formally, the multi-modal representation learning model outputs two embedding matrixes, namely location embedding matrix $\mathcal{M}_l \in \mathbb{R}^{|\mathcal{L}|\times d_w}$ and user embedding matrix $\mathcal{M}_u \in \mathbb{R}^{|\mathcal{U}|\times d_w}$, where d_w is the embedding size, $|\mathcal{L}|$ and $|\mathcal{U}|$ are the unique numbers of location/POI set and user set, respectively. Given a user trajectory $\mathcal{C}_u$, its location list as $\{l_1, l_2, \cdots, l_N\}$ is fed to location embedding layer to lookup the input representation list $\{x_1, x_2, \cdots, x_N\} \in \mathbb{R}^{N*d}$. Then the LSTM updates the hidden states at each time step t as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{16}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{17}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{18}$$

$$g_t = tanh(W_g \cdot [h_{t-1}, x_t] + b_g), \tag{19}$$

$$c_t = f_t * c_{t-1} + i_t * g_t, \tag{20}$$

$$h_t = o_t * tanh(c_t), \tag{21}$$

where x_t is the input at time t, h_{t-1} is the last output of LSTM unit, d is the hidden dimension size, σ is the sigmoid activation function, and i_t, f_t, and o_t indicate the input gate, forget gate and output gate, respectively. $W_i, W_f, W_o, W_g \in \mathbb{R}^{d*(d+d_w)}$ are gate matrix parameters, $b_i, b_f, b_o, b_g \in \mathbb{R}^d$ are bias vectors for different gates, c_t is the candidate and h_t is the output result.

Intuitively, not all locations in historical trajectory contribute equally to affect next POI selection. We bring in user attention to capture the crucial components over the location hidden vector sequences for successive POI recommendation. The user embedding vector v_u is used to enhance the influence of important location hidden states and finally we could

```
Algorithm 1 Joint training of multi-modal representation
Input: A multi-modal check-in graph $G$, embedding size $d$, iteration times $T$, negative samples $K$, hyperparameter $\beta$, learning rate $lr$.
Output: location embedding $v_l$ and user embedding $v_u$.
Initialize location and user embedding vectors $v_l = [v_c; v_s; v_t]$, $v_u$ with uniform distribution.
iter ← 1;
while iter≤ T do
    sample an edge $(l_i, l_j)$ from $E_{clcl}$, draw K negative examples and update $v_{c_i}$ and $v_{c_j}$ representations.
    sample an edge $(l_i, l_j)$ from $E_{slsl}$, draw K negative examples and update $v_{s_i}$ and $v_{s_j}$ representations.
    sample an edge $(l_i, l_j)$ from $E_{tltl}$, draw K negative examples and update $v_{t_i}$ and $v_{t_j}$ representations.
    sample an edge $(u, l)$ from $E_{ul}$, draw K negative examples and update $v_u$ and $v_c$ representations.
    sample an edge $(u_i, u_j)$ from $E_{uu}$, draw K negative examples and update $v_{u_i}$ and $v_{u_j}$ representations.
end while
```

obtain an aggregated user trajectory representation vector m for final recommendation. The process is formalized as follows:

$$a_i = \frac{exp(score(h_i, v_u))}{\sum_{k=1}^{N} exp(score(h_k, v_u))} \tag{22}$$

$$m = \sum_{i=1}^{N} a_i h_i \tag{23}$$

where a_i measures the importance of $i-$th location based on user's preference. $score(\cdot)$ is a score function which scores the importance of locations for composing trajectory representation. The score function is defined as:

$$score(h_i, v_u) = tanh(W_s[h_i, v_u] + b_s) \tag{24}$$

where $W_s \in \mathbb{R}^{2d}$ and $b_s \in \mathbb{R}^1$ are the score weight matrix and bias, respectively. The obtained user trajectory representation m will be fed to a softmax layer for generating the target POI.

$$p = softmax(W_p * m + b_p), \tag{25}$$

where $p \in \mathbb{R}^C$ is the probability distribution for all possible POIs, $W_p \in \mathbb{R}^{C \times d}$ and $b_p \in \mathbb{R}^C$ are the weight matrix and bias, respectively. Here C indicates the number of all possible POIs.

4.4. Training Algorithms

In terms of training the multi-modal representation model, the objective function 15 can be optimized with stochastic gradient descent using the techniques of edge sampling (Tang et al., 2015) and negative sampling (Mikolov et al., 2013). For each edge $(i, j) \in E$, we randomly sample multiple negative edges from a noise distribution. The sampling mechanisms can improve effectiveness of stochastic gradient descent in learning graph embeddings. In addition, we employ joint training approach to train the multi-modal representation learning model, which is described in Algorithm 1. In joint training, all types of edges are used together and then deploy edge sampling, which samples an edge for model updating at each step, where the sampling probability proportional to its weight.

5. Experiments

We conduct extensive experiments on the following aspects: (1) overall performance of our model and performance comparisons with state-of-the-art POI recommendation models, (2) performance effects of five check-in factors and effectiveness of the user attention mechanism and (3) hyper-parameter sensitivity.

5.1. Dataset Description

We choose one public dataset in (Chang et al., 2018) to evaluate our model, since it contains not only regular POI check-in sequences but also a large amount of POIs' textual comments. The statistics of this dataset are summarized in Table 1. We divide the dataset into training set, validation set and test set to conduct experimental evaluations. The most recent 20% check-ins of each user are used as test set. And the less recent 10% check-ins are used as validation set. The remaining 70% check-ins in the dataset are used as training set.

Description	Number	Description	number
check-ins	2,216,631	Average check-ins per user	28.3
POIs	13,187	Average POIs per user	15.2
Users	78,233	Average users per POI	90
Words	958,386	Average text words per POI	22.7

Table 1: The dataset statistics.

5.2. Evaluation Metrics

We adopt two metrics: Recall@k and Mean Reciprocal Rand (MRR) to evaluate model performance, which are all popularly used in ranking tasks. Recall@k is computed if the target POI is in the top-k recommended POI list. We report evaluation results on Recall@k for $k = 1$, 5 and 10. MRR represents the mean reciprocal rank of the true successive POI target in the predicted ranked lists. The high evaluation score indicates that the target POI is highly ranked in the predicted recommended list.

5.3. Baseline Algorithms

We compare the proposed model with the following baseline algorithms, which are shown as follows.

- **GRU:** utilizes the conventional GRU network and average hidden vector for successive POI recommendation.

- **LSTM:** employs a basic LSTM network and obtains the average vector to predict successive POI.

- **ST-RNN:** (Liu et al., 2016a) extends RNN model to capture temporal influence and geographical distance information between POIs.

- **STELLAR:** (Zhao et al., 2016) proposes a spatial-temporal latent ranking method to exploit interactions among user, POI and time. This model is established upon pairs of consecutive POIs to predict the next POI.

- **Geo-Teaser:** (Zhao et al., 2017) proposes a geo-temporal sequential embedding model, which exploit sequential check-in information and incorporates POI's temporal characteristics in different days.

- **CAPE:** (Chang et al., 2018) proposes a content-aware POI hierarchical embedding model, which extracts POIs' characteristics from text comments.

5.4. Parameter Setting

In our experiments, word embeddings for these methods are initialized by Glove (Pennington et al., 2014). The dimension size of embedding and hidden state d are set to 300. The weight and bias are initialize by sampling from a uniform distribution $U(-0.01, 0.01)$. The context window b is set to 5, the learning rate is set to 0.01.

5.5. Overall Performance Evaluation

Table 2 shows performance comparison results of MMR with other baseline methods. We could have following observations.

(1) **ST-RNN**, **GRU** and **LSTM** model the sequential pattern of user trajectories to predict successive POI. They achieve good performance, which demonstrates the effectiveness of POI sequential feature. **MMR** outperforms the above methods since it not only considers the sequential information but also models other POI context aspects, such as the co-occurrence, spatial connection and textual semantic relationship.

(2) **STELLAR** and **Geo-Teaser** obtains better results than above methods by further considering spatial influences among POIs. **MMR** performs better than them by further modeling user preference in a unified way.

(3) **CAPE** gains better performance compared with previous methods, which models POI co-occurrence, textual comment and geographical influence. **MMR** performs better than **CAPE**, since we model POI context and user preference jointly.

Our proposed model consistently outperforms state-of-the-art POI recommendation models due to the following considerations. Firstly, **MMR** utilizes more check-in factors, namely POI's co-occurrences, spatial/textual connections, and users' preferences. Each of them contains valuable information for POI recommendation. Secondly, **MMR** jointly trains user and location representations, where user representation is further employed as attention signals when predicting successive POI in recurrent neural network. In other words, the learned user embeddings help to enhance important locations to contribute more in final recommendation, which is aligned with user preference setting.

Model	Recall@1	Recall@5	Recall@10	MRR
GRU	0.1197	0.2207	0.2726	0.1792
LSTM	0.1207	0.2225	0.2751	0.1805
ST-RNN	0.1185	0.2142	0.2529	0.1721
STELLA	0.1308	0.2251	0.2923	0.1857
Geo-Teaser	0.1291	0.2334	0.2980	0.1850
CAPE	0.1390	0.2433	0.3079	0.1953
MMR	**0.1538**	**0.2571**	**0.3148**	**0.2097**

Table 2: The performance comparisons between MMR and other baseline methods. The results of baseline methods are retrieved from published papers. The best performances are marked in bold.

5.6. MMR Performance Analysis

After constructing multi-modal check-in graph, MMR model is able to jointly consider co-occurrence location-location influence, spatial location-location influence, textual location-location influence, user-location influence and user-user influence by optimize user and location representations to preserve second-order proximity in the graph. In this section, we study the effect of modeling each edge type. In addition, we conduct experimental evaluation to evaluate the user attention mechanism applied in recurrent neural network.

As shown in Table 3, we evaluate the following six MMR variants: w/o co-occurrence location-location model, w/o spatial location-location model, w/o textual location-location model, w/o user-location model, w/o user-user model and w/o user-attention model. For the first five variants, it is generated by ignoring the corresponding edge type. w/o user-attention model does not apply the user attention mechanism to the prediction.

Compared with complete version MMR, all variants lead to slight performance descend. These experiments demonstrate the effectiveness of modeling multi-modal check-in aspects. The results also verify that all these aspects could bring valuable information for POI recommendation. Specifically, the learned user representation is further used to guide the final recommendation in recurrent neural network as attention signals, which shows the general user preference will affect how to recommend successive POI.

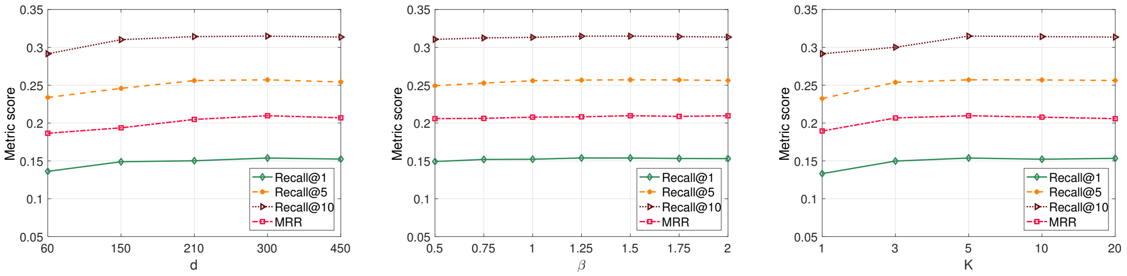
5.7. Hyper-parameter Sensitivity

In this section, we report evaluation results of hyper-parameter sensitivity. The hyper-parameters include dimension number d, control parameter β and negative samples K.

To evaluate the impact of dimension number d, we vary d from 60 to 450, the results are shown in Figure 3a. The performance is robust to the variation of d. The best performance is achieved when d is set to 300.

Variants	Recall@1	Recall@5	Recall@10	MRR
w/o co-occurrence location-location	0.1467	0.2505	0.3118	0.2026
w/o spatial location-location	0.1480	0.2533	0.3063	0.2018
w/o textual location-location	0.1505	0.2521	0.3059	0.2057
w/o user-location	0.1510	0.2499	0.3099	0.2043
w/o user-user	0.1491	0.2480	0.3116	0.2055
w/o user-attention	0.1487	0.2475	0.3074	0.2021
MMR	**0.1538**	**0.2571**	**0.3148**	**0.2097**

Table 3: The performance comparisons between MMR variants.



(a) The effect of d. (b) The effect of β. (c) The effect of K.

Figure 3: Parameter sensitivity

We vary β from 0.5 to 2 to evaluate the impacts of parameter β, which is shown in Figure 3b. When we increasing β, all metric score first slightly increases and then slightly decreases. The best performance is achieved when β is set to 1.5.

To test the impacts of negative samples K, we conduct experimental evaluation to vary K from 1 to 20, which is shown in Figure 3c. A remarkable performance improvement is observed when varying K from 1 to 3. The best performance is achieved when K is set to 5.

5.8. Case Study

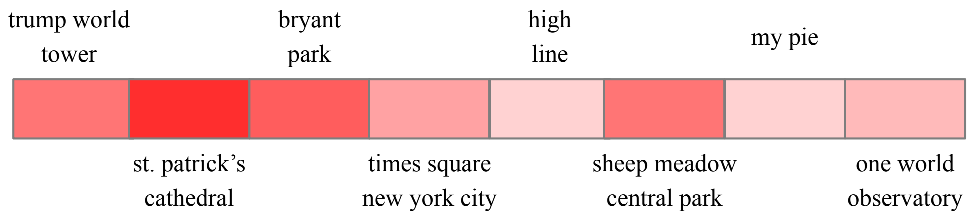
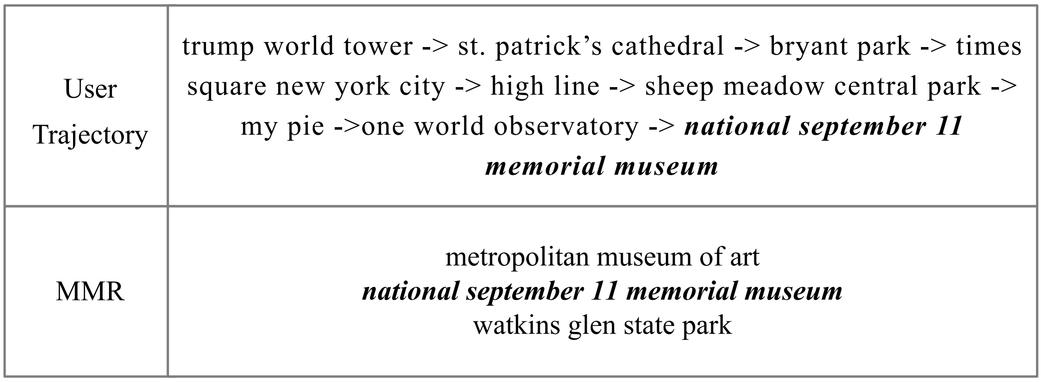


Figure 4: The user attention visualizations with a user-trajectory example.

In order to demonstrate the effect of the user attention mechanism, we visualize the attention weights of location sequences with an example shown in Figure 4. We can observe that the attention mechanism can enforce the model to pay more attentions on the important location with respect to the user's preference. For example, some POIs, such as "trump world tower" and "bryant park" have higher attention weights compared with other POIs.



User Trajectory	trump world tower -> st. patrick's cathedral -> bryant park -> times square new york city -> high line -> sheep meadow central park -> my pie ->one world observatory -> ***national september 11 memorial museum***
MMR	metropolitan museum of art ***national september 11 memorial museum*** watkins glen state park

Figure 5: POI recommendation results given a historical trajectory. The target POI is highlighted in bold and italic.

Figure 5 gives an example of recommended top-3 POIs by MMR given a historical user trajectory. We can observe that the target POI "national september 11 memorial museum" in the top three recommendations. In addition, the other two recommended POIs have the similar semantic characteristics with the target POI. From the user trajectory, we could also find the user prefers to visit tourist attraction POIs, such as parks and museums.

6. Conclusion

In this paper, we present MMR, at unified learning framework that models check-in from five perspectives, namely co-occurrence/spatial/textual location-location connections to model POI context and user-location, user-user connections to model user preference. We define a multi-modal check-in graph that encodes rich semantics from check-in records. Then we apply joint representation learning on the constructed graph to generate location and user representations. Finally, an attentional recurrent neural network is employed for final successive POI recommendation. Extensive experiments demonstrate the effectiveness of MMR on a public dataset.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61772307), and National Key R&D Program of China (2018YFB1800405, 2017YFB0801701).

References

Buru Chang, Yonggyu Park, Donghyeon Park, Seongsoon Kim, and Jaewoo Kang. Content-aware hierarchical point-of-interest embedding model for successive poi recommendation. In *IJCAI*, pages 3301–3307, 2018.

Meng Chen, Yang Liu, and Xiaohui Yu. Nlpmm: A next location predictor with markov modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 186–197. Springer, 2014.

Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2018.

Feifan Fan, Yansong Feng, and Dongyan Zhao. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, 2018.

Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Content-aware point of interest recommendation on location-based social networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Qiang Gao, Goce Trajcevski, Fan Zhou, Kunpeng Zhang, Ting Zhong, and Fengli Zhang. Trajectory-based social circle inference. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 369–378. ACM, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Dejiang Kong and Fei Wu. Hst-lstm: A hierarchical spatial-temporal long-short term memory network for location prediction. In *IJCAI*, pages 2341–2347, 2018.

Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 617–628. IEEE, 2018.

Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016a.

Xin Liu, Yong Liu, and Xiaoli Li. Exploring the context of locations for personalized location recommendations. In *IJCAI*, pages 1188–1194, 2016b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Zhu Sun, Jie Yang, Jie Zhang, Alessandro Bozzon, Yu Chen, and Chi Xu. Mrlr: Multi-level representation learning for personalized ranking in recommendation. In *IJCAI*, pages 2807–2813, 2017.

Jian Tang, Meng Qu, and Qiaozhu Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM, 2015.

Hao Wang, Huawei Shen, Wentao Ouyang, and Xueqi Cheng. Exploiting poi-specific geographical influence for point-of-interest recommendation. In *IJCAI*, pages 3877–3883, 2018.

Carl Yang, Lanxiao Bai, Chao Zhang, Quan Yuan, and Jiawei Han. Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1245–1254. ACM, 2017a.

Cheng Yang, Maosong Sun, Wayne Xin Zhao, Zhiyuan Liu, and Edward Y Chang. A neural network approach to jointly modeling social networks and mobile trajectories. *ACM Transactions on Information Systems (TOIS)*, 35(4):36, 2017b.

Di Yao, Chao Zhang, Zhihua Zhu, Jianhui Huang, and Jingping Bi. Trajectory clustering via deep representation learning. In *2017 international joint conference on neural networks (IJCNN)*, pages 3880–3887. IEEE, 2017.

Di Yao, Chao Zhang, Zhihua Zhu, Qin Hu, Zheng Wang, Jianhui Huang, and Jingping Bi. Learning deep representation for trajectory clustering. *Expert Systems*, 35(2):e12252, 2018.

Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 103–112. ACM, 2014.

Shenglin Zhao, Tong Zhao, Haiqin Yang, Michael R Lyu, and Irwin King. Stellar: spatial-temporal latent ranking for successive point-of-interest recommendation. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web companion*, pages 153–162. International World Wide Web Conferences Steering Committee, 2017.

Wayne Xin Zhao, Feifan Fan, Ji-Rong Wen, and Edward Y Chang. Joint representation learning for location-based social networks with multi-grained sequential contexts. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(2):22, 2018.