World Wide Web (2021) 24:631–655
https://doi.org/10.1007/s11280-021-00865-8

An integrated model based on deep multimodal and rank learning for point-of-interest recommendation

Jianxin Liao[1] · Tongcun Liu[1] · Hongzhi Yin[2] · Tong Chen[2] · Jingyu Wang[1] · Yulong Wang[1]

Received: 20 August 2018 / Revised: 24 June 2020 / Accepted: 25 January 2021 /
Published online: 3 March 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract
Modeling point-of-Interest (POI) for recommendations is vital in location-based social networks, yet it is a challenging task due to data sparsity and cold-start problems. Most existing approaches incorporate content features into a probabilistic matrix factorization model using unsupervised learning, which results in inaccuracy and weak robustness of recommendations when data is sparse, and the cold-start problems remain unsolved. In this paper, we propose a deep multimodal rank learning (DMRL) model that improves both the accuracy and robustness of POI recommendations. DMRL exploits temporal dynamics by allowing each user to have time-dependent preferences and captures geographical influences by introducing spatial regularization to the model. DMRL jointly learns ranking for personal preferences and supervised deep learning models to create a semantic representation of POIs from multimodal content. To make model optimization converge more rapidly while preserving high effectiveness, we develop a ranking-based dynamic sampling strategy to sample adverse or negative POIs for model training. We conduct experiments to compare our DMRL model with existing models that use different approaches using two large-scale datasets obtained from Foursquare and Yelp. The experimental results demonstrate the superiority of DMRL over the other models in creating cold-start POI recommendations and achieving excellent and highly robust results for different degrees of data sparsity.

Keywords POI recommendation · content-aware · deep multimodal networks · rank learning · cold-start

✉ Jianxin Liao
liaojx@bupt.edu.cn

✉ Tongcun Liu
tongcun.liu@gmail.com

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

[2] School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia

1 Introduction

The increasing popularity of mobile devices and advances in wireless communication technologies have resulted in the rapid development and growing use of location-based social network services (LBSNs), such as Foursquare and Yelp. LBSNs link cyberspace and the physical world, thus making it easy for users to find interesting places and to share their experiences via smart devices by posting their check-ins, finding subjective reviews, and viewing images of places. Large-scale interactive data in LBSNs is readily available, thus providing an excellent opportunity for industry and academia to gain greater insight of user preferences and behaviors [1]. There has been much recent research on point-of-Interest (POI) recommendations that provides users with accurate information about POIs (e.g., restaurants, shopping malls, or museums that they will want to visit) according to their check-in history.

Accurate modeling of user preferences is intended to provide users with high-quality POI recommendations (i.e., that are likely to appealing to the user) which are derived from the history of their interactions with POIs. This is a vital yet challenging task for the following reasons:

1) *Temporal dynamics.* User preferences are time-dependent and change continually over time [2, 3]. For example, a user may regularly arrive at the office for work on a weekday and hang out on the weekend. If the user's check-in data is directly fed into the user's preference model without time discrimination, inaccurate recommendations will be produced.
2) *Sparse data.* User check-ins are voluntary acts, and most users do not check in regularly for reasons such as lack of time or privacy concerns. Hence, the number of POIs visited by an individual user is rather small compared with the total number of POIs recognized in an LBSN. Moreover, most users visit only a limited number of POIs which are usually located within a short distance to where they live or work [4], thus further aggravating the problem of data sparsity.
3) *Cold-start*. There are cold-start POIs in LBSNs. They consist of existing POIs that have never been visited by users and new POIs that emerge during urban development. LBSNs need to recommend such POIs to targeted users to further improve user participation in LBSNs and increase revenue for POI owners. Cold-start POI recommendation is a challenging problem because there is no user–POI interaction data available [5].

The availability of huge amounts of content information (e.g., publicly available reviews and images) about POIs offers a new perspective on solutions to the above issues. Various content-aware approaches have been proposed to model the latent relations between users and POIs by leveraging reviews [6–8]. However, most of these approaches use a topic model to extract semantic features from reviews in haphazard ways that are disconnected from recommendations. They are ineffective in improving recommendation performance because they do not embed any deep understanding of the reviews. Recently, informed by the success of deep learning in creating recommendations [9–13], Zhang et al. [11] and Yin et al. [12] used supervised deep learning techniques to extract semantic features from online reviews driven by user feedback information (i.e., check-in data) to provide recommendations. The major drawback of this content-aware approach is that public reviews of POIs are aggregated into a pseudo-document, and the semantic features of the POIs are extracted from the pseudo-document without attention to the favorability of the reviews. Inaccurate recommendations are

produced because they do not fully capture the overall sentiments that are implicit in the words of a review.

Our experiments show that existing content-aware models have weak robustness against data sparsity. Recommendation accuracy decreases sharply as data becomes sparser. This is mainly because those content-aware methods regard POI recommendations as a rating prediction problem. As a result, they use probabilistic matrix factorization techniques to model latent relations between users and POIs while treating content information as a regularization term in the objective function. The inherent problem of probabilistic matrix factorization is that it cannot deal with sparse data, yet there is a lack of available check-in data. Although content information can compensate for data sparsity to some extent, the performance gain is very little when compared with the decrease in performance caused by data sparsity. Furthermore, images contained in data that provide visual descriptions of POIs have not been adequately investigated in previous studies. The process that users make decisions to visit a POI is not only influenced by community sentiment but is also sensitive to the visual attractiveness of the POI. When being modelled simultaneously, these factors can greatly improve the POI recommendation, but considering only one type of data source will lead to dissatisfaction in recommendations.

The goals of this paper are resolving the cold-start issue and improving the robustness of POI recommendations under conditions of data sparsity. We have developed a deep multimodal rank learning model (DMRL) to infer the relationships that determine different POI rankings for each user. The model is a generative probabilistic model that combines multimodal content information of POIs with a Bayesian personalized ranking (BPR) learning framework. Specifically, we establish a time-dependent user preference model and characterize a user's geographical preferences by introducing spatial regularization. This allows DMRL to capture the temporal dynamics of user behavior and synchronous geographical influences. A deep multimodal network, which links implicit feedback and semantic data, is created and integrated into BPR. This allows us to extract content and information of semantic representations having different modalities in a task-oriented and supervised way, which can improve the quality of cold-start POI recommendations. Sufficient training with both observed and implicit data ensures DMRL recommendations are robust despite data sparsity. In addition, we adopt a ranking-based adaptive sampling strategy to accelerate convergence and improve model accuracy during model optimization. We compare our model with existing recommenders using two datasets obtained from Foursquare and Yelp and demonstrate the superiority of DMRL with cold-start POI recommendations. Experiments also show that DMRL can provide excellent recommendations that are robust under different sparse data conditions.

The key contributions of the work presented in this article are as follows.1) We establish a time-dependent user preference model by exploiting the temporal dynamic of the user's behavior and geographical influences of POIs, which provides a new view to deeply understand the user preference.2) To the best of our knowledge, we are the first to obtain the semantic information of POIs from different modalities data by design a deep multimodal network, and the network was integrated into the BPR framework to link implicit feedback and semantic data in a task-oriented and supervised way.3) We develop a ranking-based dynamic sampling strategy to accelerate convergence and improve the model accuracy during model optimization.

The remainder of this paper is organized as follows: section 2 reviews related work; section 3 provides the preliminaries necessary and identifies problems for the development of the DMRL model; section 4 describes the major parts of our model; section 5 details the

experimental configuration; section 6 discusses the experimental results; the conclusion follows in section 7.

2 Related work

In this section, we discuss related work in the literature from three aspects: POI recommendation, learning ranking, and multimodal deep learning.

2.1 POI recommendation

Two paradigms govern existing methods of alleviating data sparsity and cold-start problems. The first is *context awareness* [13, 14], which exploits the inherent geographical and temporal characteristics of POIs to reduce the effects of sparse data [15–18]. Most of the context-aware methods are based on probabilistic matrix factorization [2, 19] or tensor factorization [20, 21] which considers context to be a separate dimension. However, these methods are ineffective in cold-start POI recommendation systems because context information is unavailable for them. The second paradigm is *content awareness*, which uses information about POI content to determine latent relations between users and POIs [6–8]. Using content information to infer user preferences can greatly improve the quality of recommendations [16, 22, 23]. Earlier work [6, 22] created and extended existing models by using probabilistic matrix factorization (PMF) to incorporate content features extracted using a latent Dirichlet allocation (LDA) model. Some studies [24] also modeled users' location- and attitude-relevant interests from textual content with LDA. Encouraged by the success of word2vec, models to embed users, items, and textual content in the same latent space are developed [25]. However, these content-aware methods combined content features fairly loosely and were not particularly effective in improving recommendation quality.

More recently, deep learning approaches have offered improvement in providing standard recommendations. For example, Wang et al. [9] proposed a collaborative deep learning model (CDL) that uses both deep learning for the textual content and collaborative filtering for the rating predictions. Kim et al. [10] combined a convolutional neural network (CNN) with PMF to develop a convolutional matrix factorization model (ConvMF). Zhang et al. [26] developed a collaborative knowledge-based embedding framework (CKE) to combine learning of the latent relations between users and items with semantic representations from the knowledge base in a unified framework. The success of deep learning in generating recommendations has encouraged some researchers [12] to introduce deep learning into POI recommender algorithms. For example, by combining deep representational learning with hierarchically additive representational learning, Yin et al. [12] proposed a spatially aware hierarchical collaborative deep learning model (SH-CDL). Wang et al. [27] improved the performance of the model by incorporating visual features using CNN. These studies used only one type of content information.

This study differs from previous work because we extract semantic representations of POIs from visual content and public reviews using a deep multimodal network, and combine them into a ranking learning framework to improve the quality of POI recommendations.

2.2 Machine learned ranking

Machine learned ranking (MLR) algorithms were originally developed for information retrieval tasks and have delivered state-of-the-art performance in recommender systems. There are three major categories of MLRs: pointwise, listwise, and pairwise. Pointwise approaches [28] assume that each item pair of the set to be ranked has an ordinal score, and ranking is formulated as a regression problem. This approach does not consider interdependencies among items. Listwise techniques [29] aim to directly optimize the function used to determine the ranking, but the optimization problem is complex because it has to deal with nonconvex, non-differentiable, and discontinuous functions. Pairwise methods [30, 31], which can handle sparse data, have been shown efficient and effective in ranking implicit feedback datasets by directly optimizing pairwise rankings of paired positive feedback and negative feedback.

MLR algorithms have revealed great potential in POI recommendation. For example, Feng et al. [32] used embedded metrics and ranking to model personalized check-in sequences; Li et al. [33] proposed a rank based factorization model for POI recommendations that ranks geographical characteristics; and Zhao et al. [34] built a pairwise ranking-based tensor factorization algorithm for successive POI recommendations.

2.3 Multimodal deep learning

Multimodal deep learning has progressed in many applications, partially because large-scale multimedia data is readily available. Multimodal deep learning combines data from different modalities into a conjoint representation to capture the real-world concept, or entity that the data represents [35, 36]. Ngiam et al. [35] developed a bimodal deep learning autoencoder for cross-modality shared representational learning. Zheng et al. [36] developed a multimodal emotion recognition framework that combines brain activity and eye movements. Wang et al. [37] describe an adversarial cross-modal retrieval method that searches for a cross-modal common subspace using adversarial learning.

The success of multimodal deep learning in visual computation and in information retrieval has led to studies that introduce this technique into recommender systems [38, 39]. Oramas et al. [38] combined text and audio information with user feedback data and used deep multimodal networks to address the cold-start problem in recommenders. Wang et al. [40] studied POI semantics, making use of the abundant heterogeneous user-generated content that is readily available. Zhao et al. [39] addressed the sparse data problem by creating a multimodal network for learning ranking and making recommendations.

3 Preliminaries and problem definition

For ease of representation, we first introduce some necessary preliminaries and then formulate the problems investigated in this paper.

3.1 Preliminaries

Following conventional symbol notation, we use uppercase bold letters to denote matrices (e.g., $\mathbf{Q}$), lowercase bold letters to denote row vectors (e.g., $\mathbf{q}$) and regular typeface letters to

represent scalars. We use calligraphic letters to represent sets (e.g., the user set $\mathcal{U}$), and $|\cdot|$ to denote the size of the set. The major notation used in this paper is shown in Table 1.

We make the following three definitions.

3.1.1 Definition 1 (temporal state)

A temporal state $t_k \in \mathcal{T}$ is a discrete time slot which represents the time period when a user checks in. In keeping with previous work [2], we set $|\mathcal{T}| = 2$ for $\mathcal{T}$ = {weekday, weekend}, $|\mathcal{T}| = 7$ for the seven days of a week, and $|\mathcal{T}| = 24$ for the twenty-four hours of a day. For example, if we defined $|\mathcal{T}| = 24$, then $t_k = 1$ for check-in time at "2018-04-24 00:30:00", indicating the check-in happens during the 1-h interval 0 to 1.

3.1.2 Definition 2 (implicit feedback cuboid)

Given user sets $\mathcal{U}$, POI sets $\mathcal{V}$, and temporal state sets $\mathcal{T}$, user feedback for POIs is given by $S \subseteq \mathcal{U} \times \mathcal{V} \times \mathcal{T}$. We use a cuboid $\boldsymbol{C} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{V}| \times |\mathcal{T}|}$ to store implicit user feedback. A cell c_{ijk} stores the feedback of user u_i for POI v_j during temporal state t_k. If the interactions between user u_i and POI v_j have been observed, $c_{ijk} = 1$ otherwise $c_{ikj} = 0$. Note that the value 0 in $\boldsymbol{C}$ does not mean that the users dislike the POI, but that c_{ijk} can be regarded as a mixture of negative feedback and potential interactions (i.e., users have not been aware of such POIs).

3.1.3 Definition 3 (relative preference)

For each user, we define a relative preference by an ordered tuple (u_i, v_j, t_k, v_n), meaning that the user u_i preferred POI v_j over POI v_n during time period state t_k. The set of all relative preferences, $D_S \subseteq \mathcal{U} \times \mathcal{T} \times \mathcal{V} \times \mathcal{V}$, used in our model consists of all relative preferences for all users, such that $(u_i, v_j, t_k, v_n) \in D_S : \Longleftrightarrow v_j \in \mathcal{V}^+(u_i, t_k) \wedge \mathcal{V} \setminus \mathcal{V}^+(u_i, t_k)$, where $\mathcal{V}^+(u_i, t_k) := \{u_i : (u_i, v_j, t_k) \in \boldsymbol{C}\}$ is the (observed) implicit feedback.

3.2 Problem definition

The aim of MLR, which has been widely used in recommenders, is to find a ranking function $f(\cdot)$ which will rank $\widehat{r}(\cdot)$ items for each user. The ranking function is modeled via a preference scoring function for users which is parameterized by a set of model parameters Θ.

Table 1 Key Mathematical Notations

Variable	Interpretation
$\mathcal{U}, \mathcal{V}, \mathcal{T}$	Set of users, POIs, and temporal states
$\boldsymbol{C}$	Implicit feedback cuboid of users on POIs
D_S	Set of relative preferences for all users
$\boldsymbol{p}_{i,k}$	Latent factor vector of user u_i under temporal state t_k
$\boldsymbol{q}_j$	Latent factor vector of POI v_j
$\boldsymbol{W}$	Geographical similarity matrix of POIs
$\boldsymbol{x}_j$	Crowd semantic feature vector for POI v_j
$\boldsymbol{y}_j$	Visual feature vector for POI v_j
$\boldsymbol{z}_j$	Semantic representation vector for POI v_j

In this paper, we examine recommendation ranking, which generates a ranked POI list by learning an effective ranking function $f(\cdot)$. Let $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{T}$ be sets of users, POIs, and temporal states. Given a check-in tuple $x_{ijk} = (u_i, v_j, t_k)$, we form a query q with the ith user, jth POI, and kth temporal state. Our goal is to find the top-N POIs from $\mathcal{V}$ that match the preferences of query q.

4 Proposed model

4.1 Overview

A graphical representation of the DMRL model is shown in Fig. 1. It is a generative ranking-learning framework that models temporal dynamics, geographical influences, and semantic representations of POIs; different colors represent different components: light-colored circles are observable variables, and the other circles are stochastic variables that are estimated within the model.

We project users and POIs into a K-dimensional real latent space. We assume user preferences are time-dependent to capture the temporal dynamics of user preferences. The latent vector of user u_i during temporal state t_k is denoted by $\boldsymbol{p}_{i,k}$, and $\boldsymbol{q}_j$ denotes the latent vector of POI v_j. The parameters of the preference model are denoted by Θ (i.e., $\Theta = \{\boldsymbol{p}_{i,k}, \boldsymbol{q}_j | u_i \in \mathcal{U}, t_k \in \mathcal{T}, v_j \in \mathcal{V}\}$). Our model ranks the POIs v_j given the user u_i and temporal state t_k via a preference score function $f(x_{ijk}; \Theta)$. To capture the geographical influence of user preference, geographical neighborhood relations among POIs were incorporated into $f(x_{ijk}; \Theta)$, which serves as a spatial regularization for user preferences. To overcome the cold-start POI recommendation problem in LBSNs, DMRL incorporates a semantic representation of POIs extracted from multimodal content information using a deep multimodal network. We use $\boldsymbol{z}_j$ to denote the semantic content of POI v_j and assume that the dimension of $\boldsymbol{z}_j$ is the same as the dimension of $\boldsymbol{q}_j$. The semantic representation of POIs in our model links implicit feedback preferences and semantic data. The generative process of our model is: 1) for each user $u_i \in \mathcal{U}$, and temporal state $t_k \in \mathcal{T}$, generate a latent vector $\boldsymbol{p}_{i,k} \sim N\left(0, \sigma_U^{-1} I\right)$; 2) for each POI $v_j \in \mathcal{V}$, generate a semantic representation vector $\boldsymbol{z}_j$ from the semantic information of the POI; 3) for

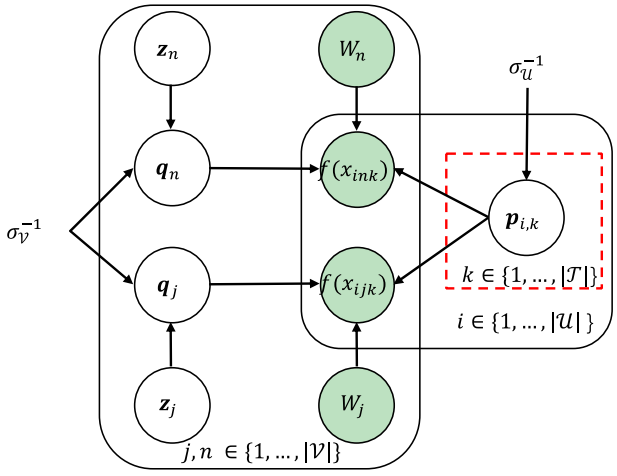

Fig. 1 Graphical representation of the DMRL model

each POI $v_j \in \mathcal{V}$, draw a latent item offset vector $\boldsymbol{\xi}_j \sim \boldsymbol{N}(0, \sigma_V^{-1} I)$ for v_j, and then set the latent vector as $\boldsymbol{q}_j = \boldsymbol{\xi}_j + \boldsymbol{z}_j$. The meaning behanding this is that the latent vector of a POI is consists of two part: content representation of the POI and the latent item offset vector, thus $\boldsymbol{q}_j \sim \boldsymbol{N}\left(\boldsymbol{z}_j, \sigma_V^{-1} I\right)$ can be considered as drawn from a gaussian distribution with $\boldsymbol{z}_j$ as mean and $\sigma_V^{-1} I$ as variance; and 4) for each tuple x_{ijk} generate a preference score using $f(x_{ijk}; \Theta)$.

Finally, following [29, 30], pairwise MLR with adaptive sampling is used for supervised learning of the latent relations between users and POIs. The following subsections give more detailed descriptions of each component of DMRL.

4.2 User preference model

The preference model is the core of DMRL. We model user preferences as the interactions between users and POIs together with geographical information. Using low rank probabilistic matrix factorization, we encode users and POIs into a K-dimensional real latent space, and we create the preference model of users by the inner product of the real-valued vectors of users and POIs. Thus, for each check-in record, x_{ijk}, the preference model is defined as:

$$f\left(x_{ijk}; \Theta\right) = \boldsymbol{p}_{i,k} \cdot \boldsymbol{q}_j{}^T \tag{1}$$

User preferences are highly sensitive to geographical proximity, and POIs located close to each other share similar user preferences [17]. To capture such geographical characteristics, we include the geographical features of POIs in the preference model by defining a $|\mathcal{V}| \times |\mathcal{V}|$ transition probabilities matrix $\boldsymbol{W}$, where w_{jl} is the probability that a neighborhood POI, v_l, influences the target POI, v_j. Following previous studies [17, 33], we set $w_{jl} = (0.5 + d(\ell_j, \ell_l))^{-1}$ if v_l is in $N(v_j)$, otherwise $w_{jl} = 0$. $N(v_j)$ is the set of geographical closest POIs to v_j. Thus, Eq. (1) with geographical influences is rewritten as:

$$f\left(x_{ijk}; \Theta\right) = \underbrace{(1-\eta)\boldsymbol{p}_{i,k} \cdot \boldsymbol{q}_j{}^T}_{user-item\ preference\ score} + \underbrace{\eta \boldsymbol{p}_{i,k} \cdot \sum_{v_l \in N(v_j)} w_{jl}\, \boldsymbol{q}_l{}^T}_{geographical\ influences\ score} \tag{2}$$

where $\eta \in [0, 1]$ is the weighting parameter used to balance the influence of geographically close items to the target POI v_j. To avoid overfitting, we enforce constraints on the latent factors of users and POIs; specifically, we set $\|\boldsymbol{p}_{i,k}\| \leq C$ and $\|\boldsymbol{q}_j\| \leq C$. The first term in Eq. (2) models the user preference score of user–POI interaction and the second term models the geographical influence, which serves as a spatial regularization of user preferences.

4.3 Representation learning

We extract the semantic information of POIs from images and public reviews by multimodal deep learning network, which is a hybrid network consisting of three subnetwork components: a deep convolutional neural network for the images, a long short-term memory autoencoder (LSTMAE) network for the public reviews, and a multimodal fusion network for learning shared semantic representation. Specifically, we use the pre-trained CNN VGG16 to extract visual representation, and use the pre-trained LSTMAE to extract semantic representation from user's reviews, then combine them via late fusion in a multimodal fusion network to create a

high-level semantic representation of POIs. Figure 2 represents the architecture of the model. The content embedding z in Fig. 1 is the output of the multimodal fusion network and serves as a middle variable in DMRL model. Notice that, during the joint learning process, only the parameters in the fusion network (the right part of the Fig. 2) was trained.

4.3.1 Visual representation

A CNN is a powerful deep network that extracts high-level visual features which can be used for for image classification and object detection. The initial layers of a CNN contain generic features (e.g., edge detectors or color blob detectors) that are useful for many applications [27, 41], so we use the pretrained CNN VGG16, which is available on ImageNet, for image classification to extract image features of POIs. VGG16 comprises 13 convolution layers, 5 maxpooling layers, 3 fully connected layers, and 1 softmax layer. Input to VGG16 is an image of size $224 \times 224 \times 3$, where 224×224 is the size of the image and 3 is the number of channels (i.e., RGB channels). We first resize each image into $224 \times 224 \times 3$ as input to VGG16, and the output is a visual representation of the image with the last two layers removed, which is a vector of dimension $d = 4096$.

4.3.2 Crowd sentiment representation

There are several accepted methods of extracting opinions from public reviews, such as Bag-of-Words (BoW), stacked denoising auto-encoders (SDAE), and CNNs. These methods usually aggregate the reviews into a pseudo-document without any distinct semantic representation of the sentiment of each review. This process cannot capture the sequential dependency between words or the volume of words in a review, and therefore leads to inaccurate results. To overcome this difficulty, we use the word-level LSTMAE model [42] to extract the sentiment expressed in each review, and then the overall sentiment expressed for a POI is obtained by accumulating the opinions from all the reviews using max pooling technology.

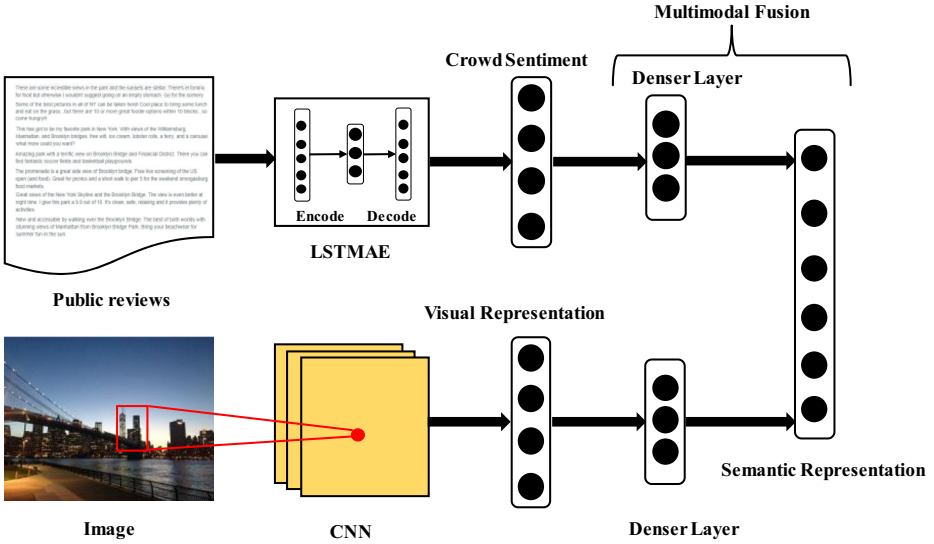

Fig. 2 Illustration of multimodal deep learning network

LSTMAE consists of two recurrent neural networks: an encoder LSTM and a decoder LSTM. The encoder LSTM receives input sequences and produces a set of vectors, and the decoder LSTM receives the latent vectors and decodes them into the original input sequences. Figure 3 shows the schematic of LSTMAE.

For each review, all the words are embedded into a high-dimension matrix via a pre-trained word embedding model, such as Glove. The review is then represented as a two-dimensional matrix, where the rows of the matrix are the number of words in the review and the columns are the length of the embedded words vector. By training the LSTMAE network, we can generate the representation of all opinions summarized for each one of the POIs.

4.3.3 Multimodal fusion

There are several approaches to multimodal learning described in the literature [12], such as early fusion and late fusion strategies. We extract visual features and public opinions separately and combine them via late fusion in a multimodal network. We normalize them and input them to a forward neural network (a simple multilayer perceptron, MLP). Each feature vector is connected to an isolated dense layer with a scaled exponential linear unit (SELU) activation function, then connected to the output layer. The reason is that the isolated dense layers help the network learn a nonlinear representation for each separate modality [38]. The model is regularized by applying dropout with an empirically determined factor of 50% to each output layer. We use $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$ to denote the images and opinions for POI $\boldsymbol{v}_j$, and define the multimodal fusion network as:

$$\boldsymbol{z}_j = MF\left(\boldsymbol{x}_j, \boldsymbol{y}_j; \boldsymbol{\Psi}\right) \tag{3}$$

The output layer of the network is the semantic representation of the jth POI, $MF(\cdot)$ denotes the fusion architecture, and $\boldsymbol{\Psi}$ represents the parameters that will be estimated in model training.

4.4 Model optimization

Pairwise learning was used to learn the parameters of DMRL. The intention is to discriminate each user with the observed POIs, $\mathcal{V}^+(u_i, t_k)$, and the unvisited or uncommented POIs, $\mathcal{V}\backslash\mathcal{V}^+(u_i, t_k)$. A POI v_j is preferred over a POI v_n by user u_i only if v_j but not v_n was visited by user u_i. We use a sampling strategy to select a negative POI v_n, one that has been badly reviewed, such that the pairwise tuple (u_i, v_j, v_n, t_k) is sufficiently informative for the current values of the model

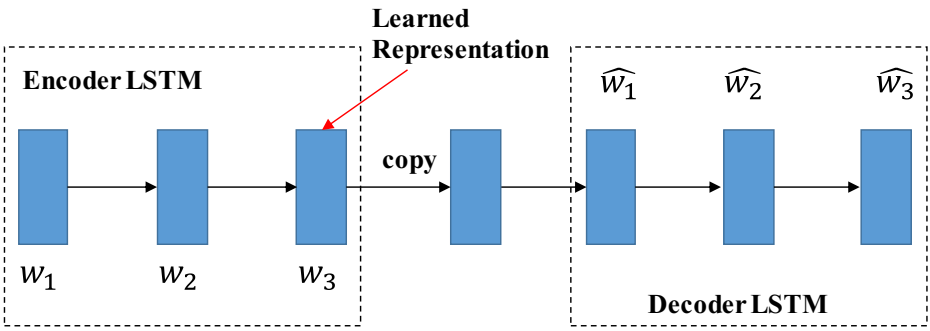

Fig. 3 Illustration of the LSTMAE network

parameters $\widehat{\Phi}$ to update their gradients [31, 43]. Thus, we adopt an adaptive sampling strategy, which considers both the context of the observation and the current values of the model parameters $\widehat{\Phi}$ during sampling. The underlying intuition is that when a negative item v_n for a given observation x_{ijk} is sampled, the closer v_n is ordered to the top and thus the more informative it becomes. This procedure is also known as a *ranking-based sampling strategy*. By using a geometric distribution, following [31], the adaptive negative sampling distribution is:

$$P(v_n|u_i,t_k)\propto exp\left(-\frac{\widehat{r}(v_n|u_i,t_k)}{\lambda}\right) \tag{4}$$

where $\widehat{r}(v_n|u_i,t_k)$ is the rank of item v_n based on the current preference score $f(x_{ikn};\Theta)$, and λ is a hyper-parameter which tunes the probability density.

Given the pairwise relation of the positive POI v_j and the sampled negative POI v_n, the link between the pair and the ranking function $f(\cdot)$ is established by:

$$p\left(j>_{i,k}n|\Theta\right):=\sigma\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right) \tag{5}$$

where $\sigma(x) = 1/(1 + exp(-x))$. Using a bayesian probabilistic generative process, the posterior probability of the personal preferences and the POI's latent preferences is formulated as:

$$\begin{aligned} p(\Phi|j>n;u_i,t_k)\propto & p(j>n;u_i,t_k|\Phi)p(\Phi)= \\ & \prod_{(i,j,n,k)\in D_S}\sigma\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)\times\prod_i\prod_k N\left(\boldsymbol{p}_{i,k}|0,\sigma_U^{-1}I\right)\times\prod_j N\left(\boldsymbol{q}_j|\boldsymbol{z}_j,\sigma_V^{-1}I\right) \\ & \times\prod_n N\left(\boldsymbol{q}_n|\boldsymbol{z}_n,\sigma_V^{-1}I\right)\times\prod_m^{|\boldsymbol{\Psi}|}N\left(\Psi_m|0,\sigma_{\boldsymbol{\Psi}}^{-1}I\right) \end{aligned} \tag{6}$$

where: $\Phi = \{\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{\Psi}\}$; $\boldsymbol{P}$ and $\boldsymbol{Q}$ are formed by $\boldsymbol{p}_{i,\,k}$ and $\boldsymbol{q}_j$; and $\boldsymbol{z}_j$ links implicit feedback preference and semantic information (i.e., textual data and visual data) of POIs. To optimize these parameters, we minimize the negative log probability:

$$\begin{aligned} \mathcal{L} = & -\sum_{(i,j,n,k)\in D_S}\ln\left(\sigma\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)\right)+\frac{\sigma_U}{2}\sum_i\sum_k\left\|\boldsymbol{p}_{i,k}\right\|^2+\frac{\sigma_V}{2}\sum_j\left\|\boldsymbol{q}_j-\boldsymbol{z}_j\right\|^2 \\ & +\frac{\sigma_V}{2}\sum_n\|\boldsymbol{q}_n-\boldsymbol{z}_n\|^2+\frac{\sigma_\Psi}{2}\|\boldsymbol{\Psi}\|^2 \end{aligned} \tag{7}$$

where $\|\boldsymbol{p}_{i,\,k}\|^2$, $\|\boldsymbol{q}_j-\boldsymbol{z}_j\|^2$, $\|\boldsymbol{q}_n-\boldsymbol{z}_n\|^2$ and $\|\boldsymbol{\Psi}\|^2$ are serve as regularized L_2 term in the loss function.

It is computationally expensive to directly minimize Eq. (7) due to summing the huge pairwise D_S. Thus a stochastic gradient descent (SGD) algorithm was used to estimate the latent factors $\boldsymbol{p}_{i,\,k}$, $\boldsymbol{q}_j$, and $\boldsymbol{q}_n$, by fixing parameter $\boldsymbol{\Psi}$, it becomes:

$$\boldsymbol{p}_{i,k}=\boldsymbol{p}_{i,k}+\alpha\left(\frac{e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}{1+e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}\left((1-\eta)\left(\boldsymbol{q}_j-\boldsymbol{q}_n\right)+\eta\left(\sum_{v_{l^+}\in N\left(v_j\right)}w_{jl^+}\boldsymbol{q}_{l^+}{}^T-\sum_{v_{l^-}\in N\left(v_n\right)}w_{nl^-}\boldsymbol{q}_{l^-}{}^T\right)\right)-\sigma_U\boldsymbol{p}_{i,k}\right) \tag{8}$$

$$\boldsymbol{q}_j = \boldsymbol{q}_j + \alpha\left(\frac{e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}{1+e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}(1-\eta)\boldsymbol{p}_{i,k}-\sigma_V\left(\boldsymbol{q}_j-\boldsymbol{z}_j\right)\right) \tag{9}$$

$$\boldsymbol{q}_n = \boldsymbol{q}_n + \alpha\left(-\frac{e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}{1+e^{-\left(f\left(x_{ijk};\Theta\right)-f\left(x_{ink};\Theta\right)\right)}}(1-\eta)\boldsymbol{p}_{i,k}-\sigma_V\left(\boldsymbol{q}_n-\boldsymbol{z}_n\right)\right) \tag{10}$$

since $\boldsymbol{\Psi}$ consists of many parameters and is closely related to the nonlinear activation function in $MF(\cdot)$. As described in [10], the loss function $\mathcal{L}$ can be treated as a squared error function with regularized L_2 terms as follows when $\boldsymbol{P}$ and $\boldsymbol{Q}$ are made constant:

$$\epsilon(\boldsymbol{\Psi}) = \frac{\sigma_V}{2}\sum_j\left\|\boldsymbol{q}_j-\boldsymbol{z}_j\right\|^2 + \frac{\sigma_V}{2}\sum_n\left\|\boldsymbol{q}_n-\boldsymbol{z}_n\right\|^2 + \frac{\sigma_W}{2}\sum_m^{|\Psi|}\left\|\boldsymbol{\Psi}_m\right\|^2 \tag{11}$$

Then, inspired by the work [10], we use the back propagation algorithm to optimize $\boldsymbol{\Psi}$.

The overall optimization process ($\boldsymbol{P},\boldsymbol{Q}$ and $\boldsymbol{\Psi}$ are alternatively updated) is repeated until convergence. Notice that, the content representation $\boldsymbol{z}_j$ will constraint the learning of POI's latent vector $\boldsymbol{q}_j$ by eq. (9) and eq. (10). Besides, the loss between $\boldsymbol{q}_j$ and $\boldsymbol{z}_j$ was used to supervise the parameters learning of multimodal fusion network by eq. (11), thus $\boldsymbol{z}_j$ was dynamically changed with the learning of the multimodal fusion network.

4.5 Fast learning scheme

The adaptive sampling process is: 1) sample a rank r from the geometric distribution $P(v_n|u_i, t_k)$; and 2) rank all items using the ranking function and return the item v_n which is ranked in the rth position: i.e., $\widehat{r}(v_n|u_i,t_k) = r$ or $j = \widehat{r}^{-1}(v_n|u_i,t_k)$. The second step has to compute $f(x_{ijk};\Theta)$ for all POIs, and then sort them by the estimated scores. The complexity for the sampling is $O\left(|\mathcal{V}|\cdot T_{pre} + |\mathcal{V}|\cdot log|\mathcal{V}|\right)$, T_{pre} is the time for predicting a score. In practice, it is not feasible to perform this calculation.

To resolve this issue, we use an approximate sampling algorithm for fast learning, implemented in the following steps: 1) sample a rank r from the geometric distribution $P(v_n|u_i, t_k)$; 2) draw a dimension f from $P(f|\boldsymbol{p}_{i,k}) \propto \boldsymbol{p}_{i,k,f}\cdot\sigma_f$, where $\boldsymbol{p}_{i,k,f}$ denotes the value of the latent user vector $\boldsymbol{p}_{i,k}$ on dimension f and $\sigma_f = \mathrm{Var}(\boldsymbol{p}_{\cdot,\cdot,f})$; 3) sort all items in descending order according to the values of the latent vector of dimension f, denoted as $\widehat{r}^{-1}(\cdot|f)$; and 4) return the negative item v_n in position r. Note that it is still computationally time-consuming to calculate $\widehat{r}^{-1}(\cdot|f)$, so we precompute the ranking of POIs in $\mathcal{V}$ for each of the K dimensions every few iterations instead of at each iteration because of the observation that after a single iteration, the ranking $\widehat{r}^{-1}(\cdot|f)$ changes in a small scale. We precompute the K ranking lists of POIs every $|\mathcal{V}|\cdot\log|\mathcal{V}|$ iterations because this has been shown to produce a better performance [44]. The sampling algorithm has an expected runtime of $O(K)$ for drawing a negative item, which is the same as for a single iteration. The detailed pseudocode implementation of the DMRL optimization is shown in Algorithm 1.

5 Experimental configuration

5.1 Data description

To demonstrate the effectiveness of the model, we make use of two publicly-available datasets from Foursquare and Yelp for empirical studies. These two datasets contain abundant check-in data and have been widely used in previous work.

5.1.1 Foursquare

We use the Foursquare dataset provided by [45]. This dataset includes long-term (about 10 months) check-in data for New York city collected by Foursquare from 2012 to 04-12 to 2013-02-16, which contains 1083 users and 38,333 POI locations with 227,428 check-ins. The dataset did not include any of the review data or image data needed in this study, so we crawled such data using the Foursquare API.[1] The sparsity is 99.45% in this dataset.

5.1.2 Yelp

The Yelp challenge dataset round 11 contains 1,300,000 users and 1,200,000 businesses from 11 cities across 4 countries. Each check-in record is stored as user-ID, item-ID, item-location, item-review, item-image and check-in date. We selected check-ins in Las Vegas for our experiment by extracting users who have at least 20 reviews. We resulted in 8439 users, 20,605 POIs, and 393,329 reviews in the final dataset. The data sparsity is 99.77%.

5.2 Evaluation metrics

Given a collection of check-in records S generated by a user, we first order these records according to their timestamps. Then we use the first x% check-ins for the training set $S_{training}$ and the rest for the testing set S_{test}. The variable x was set to 10, 20, 40, and 80 in our experiment. In the training dataset with 80% of the check-ins, we choose the last 10% as the validation data to tune the hyperparameters of our model, such as learning rate, regularization parameters, count of latent factors, geometric distribution, and the number of POI neighborhoods.

The aim of this work is to find out the top-N that users may be interested in. We use two common ranking evaluation metrics, Accuracy@N and mean reciprocal rank (MRR), to evaluate the quality of the model. Those two metrics have been widely used [11, 43, 46]. For each user behavior $x_{ikj} \in S_{test}$, we proceed as follows.

We first choose 1000 POIs located around the POI v_j according to the shortest distance and have never been visited by user u_i during time interval t_k, thus forming 1000 negative examples. As noted above, only those POIs geographically close to v_j are its competitors. We then compute the score for x_{ikj} and the 1000 negative examples using the preference score function in Eq. (2) in order to form a ranked list by ordering these POIs according to their scores. We next form a top-N recommendation list by picking the top N ranked items from the list. If $\hat{r}(v_j) \leq N$, we have a hit; otherwise, we have a miss. Finally, the overall Accuracy@N is defined by averaging over all test cases:

[1] https://developer.foursquare.com/

$$Accuracy@N = \frac{\#hit@N}{|S_{test}|} \tag{12}$$

where $\#hit\ @\ N$ denotes the number of hits in the test set, and $|S_{test}|$ is the number of all test cases.

The mean reciprocal rank (MRR) evaluates a ranking task that produces a list of responses to a query that is ordered by probability of correctness. The reciprocal rank of a positive response over all negative responses is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of all results for a query, and it is defined as:

$$MRR = \frac{1}{|S_{test}|} \sum_{x_{ikj} \in S_{test}} \frac{1}{rank\left(x_{ikj}\right)} \tag{13}$$

where $rank(x_{ikj})$ is the position of POI v_j in the overall negative responses.

5.3 Comparative approaches

We compared DMRL with the following state-of-the-art recommendation models, which have been used as comparisons in many studies:

1) **CTR**. Collaborative topic regression [47] is the first instance of leveraging textual information for recommendations. In this model, LDA is used to extract content topics from an item, and to integrate them into a collaborative filtering model.
2) **CDL**. Collaborative deep learning [9] is the basis of the first deep learning recommendation model. In this model, SDAEs were used to learn the hidden feature vectors from textual information about items, which were then incorporated into PMF.
3) **RankGeoFM**. A ranking based geographical factorization model [33] that represents a state-of-the-art approach for POI recommendations. It models user preferences by incorporating geographical information and extends the ordered weighted pairwise classification (OWPC) function to rank POIs with different visited frequencies. RankGeoFM has been shown to outperform most of the classical models by [4], so we compare DMRL only with this model.
4) **VBPR.** Visual Bayesian personalized ranking [41] of implicit feedback make use of visual features extracted from item images using a pre-trained deep network to learn visual user preferences.
5) **ConvMF**. Convolutional matrix factorization [10] is a context-aware recommendation model that combines CNN with PMF. This model uses a CNN to extract contextual features of items, which is the main difference between it and CDL or CTR.

We did not compare our model with BPR [30] or PMF [48] because they have been shown to perform worse than the aforementioned comparative approaches [41, 47].

5.4 Implementation notes

We first removed punctuation, numbers, stop words, and words of less than two characters from reviews since these words usually don't have discriminative meanings. We then removed suffixes from the remaining words using the Porter stemmer. We removed reviews of less than

10 words since they cannot provide adequate semantic representation of the POIs. For image data, we randomly chose one image for each POI as visual data for the POI and resized those images into $3 \times 224 \times 224$ tensor format in the RGB color space.

We trained our model on a computer server equipped with 4 high-performance NVIDIA GPUs, each having 12 GB video memory. It is computationally time-consuming to train the DMRL model on two datasets using a single GPU, so we developed a parallel implementation of DMRL to use the full power of the GPUs on the server. Our model was implemented using python with the Keras deep learning library. We used the pre-trained VGG16 in Keras to extract visual features.

There are several hyperparameters in our model, and we performed 10-fold cross-validation to find the optimal parameters. We used a conventional grid search algorithm to obtain the optimal hyperparameters for the validation data ($\sigma_U = 0.1$, $\sigma_V = 1$, $\sigma_W = 0.01$, and $C = 1$). The sensitivity analysis for some important hyperparameters (the dimension of latent factors K, λ for the geographical distribution, the geographical weighting parameter η, and the number of neighborhoods $\#NN$) is given in section 6.3. The temporal state for weekday–weekend patterns was defined as $t = [1 : T]$ with $T = 2$, which gave the best performance in our experiment. We omitted the experimental results because of space constraints.

In the comparisons with other models, we used the source codes provided by the various authors. The optimal hyperparameters for other models were obtained by a grid search algorithm using our datasets. For CTR, we set K=50, $\lambda_u = 1$, $\lambda_v = 10$, $a = 1$, $b = 0$. For CDL, we set $a = 1$, $b = 0$, K=50, $\lambda_u = 100$, $\lambda_v = 10$. For RankGeoFM, we set $K = 32$, $\alpha = 0.7$, $C = 1$, $\varepsilon = 0.3$, $\#NN = 100$. For VBPR, we set $K = 50$, $\lambda_u = \lambda_p = \lambda_i = 1$, $\lambda_j = 1$, $\lambda_b = \lambda_e = 0$. For ConvMF, we set $K = 50$, $\lambda_u = 1$, $\lambda_v = 100$.

6 Results and discussion

6.1 Recommendation effectiveness

In this section, we assess the effectiveness of our DMRL model in resolving the cold-start POI recommendation problem and its robustness under different data sparsity settings.

6.1.1 Results on cold-start item recommendation

This experiment examines the accuracy of DMRL in producing cold-start POI recommendations. In keeping with [11, 15], we first identify POIs with less than 5 check-ins in our two datasets as cold-start items and then select users with at least one cold-start POI check-in as test users. For each test user, we select their check-in records that are associated with cold-start items as the test set and the remaining check-ins as the training set. Our aim is to test whether the identified cold-start POIs in the test set can be accurately recommended to the likely users by appearing in the top-N results.

We tested our DMRL model only against VBPR, CTR, CDL, ConvMF, and DMRL on the cold-start POI recommendations since they are able to recommend cold-start POIs. We use Accuracy@N and MRR to evaluate recommendation performance, varying N in {5, 10, 15, 20, 25, 30}. We conducted the experiment 10 times, and the average values are reported as the final results for the two datasets. The

values of Accuracy@N for the comparisons using the Foursquare and Yelp datasets are given in Fig. 4, which shows that the DMRL model significantly outperforms the other models and is more accurate than the second best model by 1.54%–1.28% on the Foursquare dataset and 1.78%–1.16% more accurate on the Yelp dataset. When DMRL is compared with ConvMF, CDL, CTR, and VBPR, Accuracy@10 shows that DMRL gives improvements of 101.78%, 19.74%, 152.22%, and 38.12% on the Foursquare dataset and 53.89%, 113.51%, 166.89%, and 444.83% on the Yelp dataset. Table 2 shows the MRR results for DMRL and other models. The results show that DMRL significantly outperforms other models; for example, the improvements over ConvMF (the second best model) are 34.8% (Foursquare) and 47.08% (Yelp).

Other important observations can be made from the results. First, DMRL achieves much higher recommendation accuracy than the state-of-the-art models on cold-start POI recommendation. This demonstrates that exploiting visual information and crowd sentiment via the deep multimodal network to model high-level semantics of POIs can significantly enhance the quality of POI recommendations [24, 27]. Images and public reviews of POIs can improve POI recommendations. Considering only one of the two will decrease the accuracy of the recommendations because user decision-making processes are highly sensitive to crowd sentiment and visual attractiveness. The deep learning multimodal network was trained under the influence of the interactions between users and POIs; thus the learned semantic representation of POIs further improves the recommendations because they are task-guided. Second, VBPR and CDL perform better than ConvMF on the Foursquare dataset but worse on the Yelp dataset. This is because the Foursquare data is denser than the Yelp data and there are more reviews and images in Foursquare than Yelp. Thus, model training for VBPR and CDL is more complete for Foursquare than for Yelp. CDL performs better than ConvMF on Foursquare data because ConvMF underfits due to the number of short reviews in Foursquare, while CDL avoids this problem by making use of word frequency for training. Third, ConvMF outperforms other alternative models using the Yelp dataset because the average review length in Yelp is greater than in Foursquare. ConvMF uses CNN, which can better understand semantics than SDAE, to extract textual features, which demonstrates the benefits of using CNN rather than SDAE on dense data.

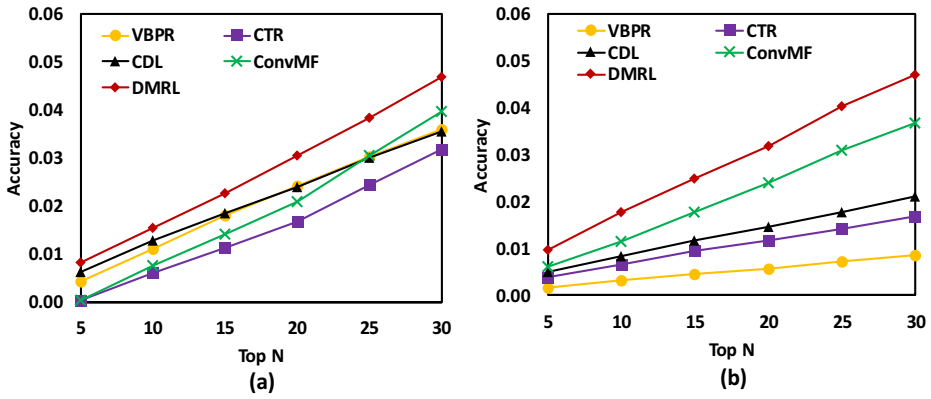

Fig. 4 Recommendation performance for cold-start POIs

Table 2 MRR in cold-start POI recommendation

Dataset	VBPR	CTR	CDL	ConvMF	DMRL
Foursquare	0.0086	0.0070	0.0085	0.0087	**0.0117**
Yelp	0.0041	0.0051	0.0064	0.0083	**0.0123**

6.1.2 Results under different sparse data settings

This section assesses the accuracy and robustness of DMRL for different data sparsity settings (i.e., 10%, 20%, 40%, and 80%). Table 3 and Table 4 show the results for Accuracy@5, Accuracy@10, and MRR using the Foursquare and Yelp datasets. Figure 5 shows the improvement in performance for Accuracy@5, Accuracy@10, and MRR compared to the best alternative models using the two datasets. DMRL significantly outperforms the alternatives for both datasets under different data sparsity setting. For example, the average improvements, in terms of Accuracy@10, are 18.3% (Foursquare) and 9.3% (Yelp). The results show the superiority of exploiting and integrating visual data and crowd sentiments for enhancing POI recommendations. Reasons for better Accuracy@10 and MRR values for DMRL are: 1) DMRL was designed for rank top-N POIs for each user by modeling temporal dynamics and geographical influences in a unified framework; and 2) visualization and crowd sentiment were exploited using a deep multimodal network and incorporated into BPR guided by user feedback information, thus further alleviating the sparse data problem.

Other important observations are also drawn from the results. First, DMRL and VBPR both give better and more robust recommendations under different data sparsity settings than the other models. This is because DMRL and VBPR use BPR, which makes use of observed and implicit data for training [30, 41], thus reducing the effects of sparse data. BPR is more robust than PMF when coping with the data sparsity problem [30]. Content information was incorporated into BPR through regularization, which can further alleviate the sparsity problem. Second, CDL and CTR perform worse under the 10% and 80% data sparsity settings and perform better under the 20% and 40% sparse data settings. This is because CDL and CTR are both PMF models, and PMF has difficulty in handling data sparsity. Although content information reduces the effects of data sparsity, the increase in performance is very limited in comparison to the decreased performance under the extreme data sparsity settings. For a denser dataset, overfitting may occur when a lot of content information is available. Third, ConvMF performs better under the 10% and 20% data sparsity settings but performs worse under the 40% and 80% settings. This is because ConvMF uses a CNN to extract content features, which is more effective when tackling a sparse dataset. Finally, RankGeoMF performs worse under the 10% data sparsity setting but better under the 80% setting because RankGeoMF was developed to rank POIs with different check-in frequencies.

6.2 Parameter tuning and sensitivity analysis

In this experiment, we investigated the sensitivity of hyperparameters (the geographical weighting parameter, number of neighborhoods, the dimension of latent factors, and the geographical distribution) in our DMRL model using the Foursquare and Yelp datasets.

The geographical weighting parameter η balances the influence of geographically close items to the target POI. To study its impact, we tested the performance of DMRL by varying η

Table 3 Recommendation results using the Foursquare dataset with four data sparsity settings

Methods	Accuracy@5				Accuracy@10				MRR			
	10%	20%	40%	80%	10%	20%	40%	80%	10%	20%	40%	80%
RankGeoMF	0.0366	0.0426	0.0552	0.0745	0.0591	0.0674	0.0852	0.1011	0.0300	0.0354	0.0460	0.0612
VBPR	0.0868	0.0836	0.0868	0.0959	0.1104	0.1207	0.1229	0.1208	0.0710	0.0707	0.0707	0.0799
CTR	0.0357	0.0501	0.0501	0.0407	0.0564	0.0751	0.0736	0.0644	0.0304	0.0372	0.0371	0.0331
ConvMF	0.0738	0.0381	0.0324	0.0245	0.1096	0.0751	0.0576	0.0448	0.0559	0.0325	0.0284	0.0246
CDL	0.0446	0.0626	0.0626	0.0509	0.0705	0.0938	0.0920	0.0805	0.0380	0.0465	0.0463	0.0414
DMRL	**0.0923**	**0.0930**	**0.0964**	**0.0998**	**0.1286**	**0.1389**	**0.1430**	**0.1511**	**0.0748**	**0.0752**	**0.0782**	**0.0826**

Table 4 Recommendation results using the Yelp dataset with four data sparsity settings

Methods	Accuracy@5				Accuracy@10				MRR			
	10%	20%	40%	80%	10%	20%	40%	80%	10%	20%	40%	80%
RankGeoMF	0.0130	0.0241	0.0417	0.0488	0.0209	0.0383	0.0692	0.0807	0.0131	0.0207	0.0326	0.0422
VBPR	0.0691	0.0714	0.0729	0.0690	0.1185	0.1235	0.1281	0.1186	0.0591	0.0602	0.0601	0.0576
CTR	0.0332	0.0610	0.0594	0.0546	0.0584	0.0998	0.1079	0.0974	0.0288	0.0472	0.0495	0.0452
ConvMF	0.0678	0.0653	0.0549	0.0416	0.1162	0.1159	0.1016	0.0769	0.0547	0.0527	0.0456	0.0360
CDL	0.0415	0.0762	0.0743	0.0683	0.0730	0.1247	0.1349	0.1217	0.0360	0.0591	0.0619	0.0564
DMRL	**0.0808**	**0.0826**	**0.0834**	**0.0866**	**0.1319**	**0.1376**	**0.1433**	**0.1453**	**0.0644**	**0.0668**	**0.0674**	**0.0676**

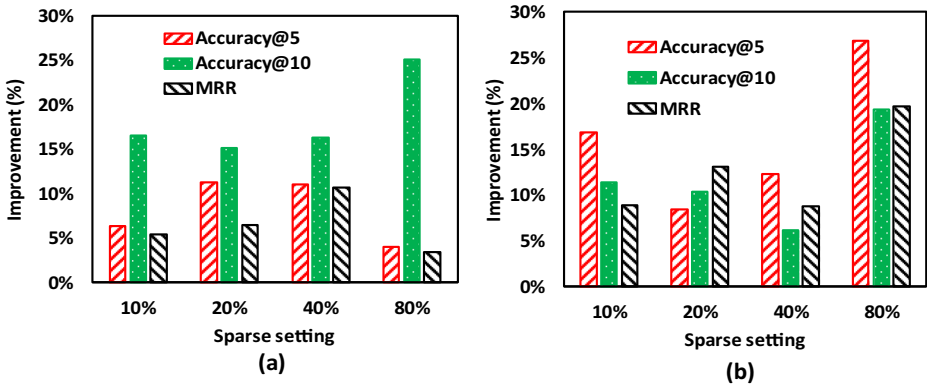

Fig. 5 Improvement for Accuracy@5, Accuracy@10, and MRR under different density setting

from 0.0 to 0.9 in increments of 0.1. The results for Accuracy@10 are shown in Fig. 6(a). When η is set to 0, it means that no neighborhood POIs are taken into account, as is also the case in Fig. 6(b). dcrease when η is larger than 0.5 (Foursquare) and 0.2 (Yelp). This suggests that η can achieve a trade-off between user preferences on the target POIs and the nearest

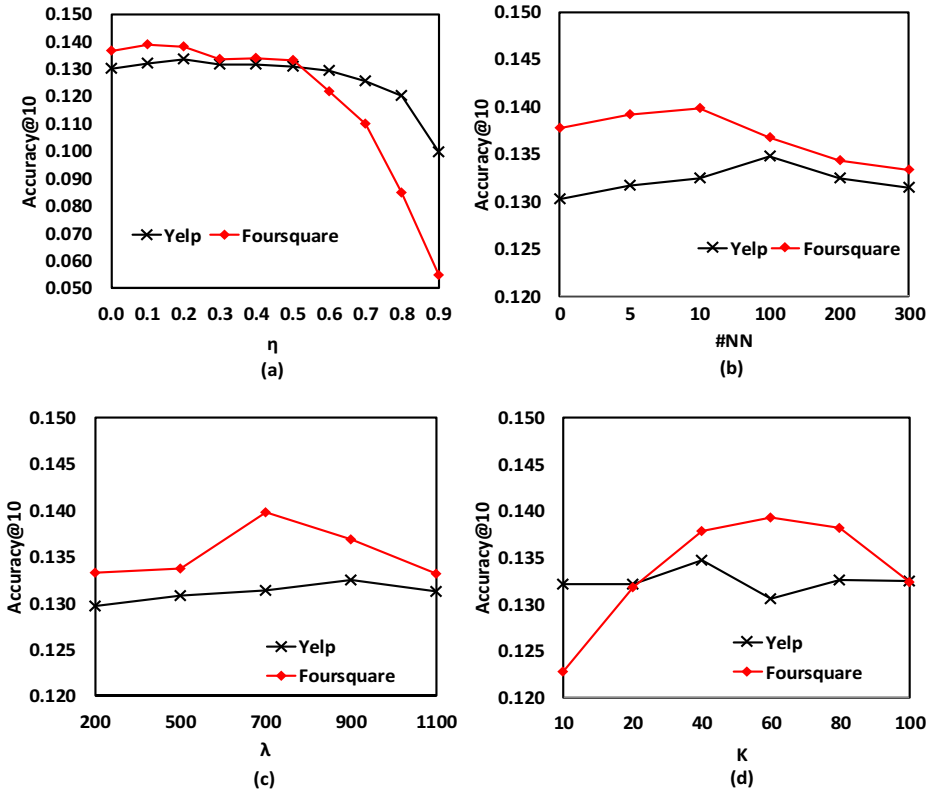

Fig. 6 Impact of the hyperparameters in DMRL

neighboring POIs, however the recommendation accuracy will decrease significantly if only one of them is considered [17, 33].

We tested the performance of DMRL by varying the number of neighboring POIs, *#NN*, from 0 to 300 to further examine the effect on recommendation accuracy of the quantity of neighboring POIs. Figure 6(b) shows that the recommendation accuracy of DMRL first increases then slightly decreases as *#NN* continues to increase. The best results are given when *#NN* is set to 10 (Foursquare) and 100 (Yelp). The reason for the early increase is that the growth in the number of neighboring POIs will increase the effect of geographical influences on user preferences. Recommendation accuracy begins to decrease slightly as *#NN* continues to increase because the increased number of neighboring POIs will increase the number of geographically unrelated POIs.

Another important hyperparameter in DMRL is λ for the geometric distribution in Eq. 5. Figure 6(c) shows the Accuracy@10 values for DMRL when λ is varied in {200, 500, 700, 900, 1100}. The results show that the recommendation accuracy of DMRL first increases as λ increases and then decreases when λ is greater than 700 (Foursquare) and 900 (Yelp).

We also investigated the effect of the dimension K on recommendation accuracy. Figure 6(d) shows the Accuracy@10 values. The performances of DMRL first improved sharply as K increased, and then decreased as K continued to increase for the Foursquare dataset. However, for the Yelp dataset, the performance of DMRL first increased slightly as K increased and then slightly decreased as K continued to increase. The hyperparameter K represents model complexity. When K is small, DMRL cannot describe user preferences. However, when K exceeds some threshold, the model is complex enough to handle the data with some accuracy. At this point, increasing K will undoubtedly improve model performance, but it will also increase the time taken for model training, leading to diminishing returns. The best results were achieved when $K = 40$ (Yelp) and $K = 60$ (Foursquare).

6.3 Analysis of different sampling strategies

We also compared the fast learning technique with the uniform sampling strategy. We use the term *DMRL-U* to denote the DMRL model using uniform sampling. Figure 7 shows the Accuracy@10 results for the two datasets as the number of training iterations increases. The results show that the fast learning technique using the ranking-based sampling strategy

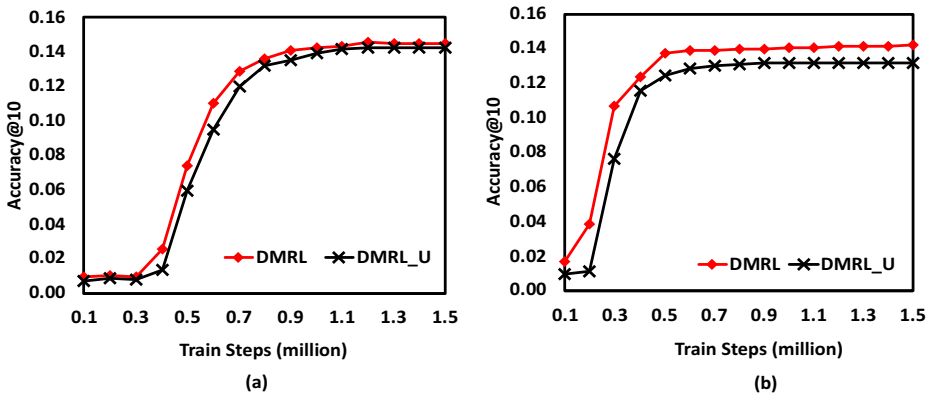

Fig. 7 Convergence analysis on Foursuare and Yelp data in terms of Accuracy@10

significantly outperforms the uniform sampling strategy. DMRL shows better results with fewer training iterations. The reason for such a significant improvement is that the adaptive sampling strategy considers both the context of the observation and the current values of the model parameters during sampling, making it more capable of sampling more informative negative POIs in each training iteration than uniform sampling.

7 Conclusion

In this work, we aimed to solve the problem of cold-start POI recommendations and the data sparsity problem by developing a deep multimodal rank learning (DMRL) model, which is a generative probabilistic model that incorporates multimodal content information into a Bayesian personalized ranking-learning framework (BPR). We exploit temporal dynamics by allowing each user to have time-dependent preferences, and we capture geographical influences by introducing spatial regularization. Our model builds relationships between implicit feedback and the semantic representation of POIs by supervised learning of ranking in a deep multimodal network. To improve the speed of model optimization and improve model accuracy, we adopted a ranking-based dynamic sampling strategy to sample negative POIs (adverse reviews or unreviewed POIs). We conducted experiments using two large-scale datasets obtained from Foursquare and Yelp, and our DMRL model significantly outperforms other state-of-the-art models in terms of accuracy and MRR. The improvements for cold-start POI recommendations offers more real-life practicality compared with the best alternative approaches. In terms of Accuracy@10, the improvements are 19.74% for the Foursquare dataset and 53.89% for the Yelp dataset. By parameterizing and varying the proportion of sparse data, the experiments showed that our DMRL model also gives better and more robust recommendations. The average improvements on Accuracy@10 are 18.3% for the Foursquare dataset and 9.3% for the Yelp dataset. In future investigation, we will incorporate the spatial dynamics of user preferences for out-of-town recommendations, and extend our DMRL model using streaming recommender system techniques for online recommendations.

Acknowledgments This work was supported in part by the National Key R&D Program of China 2020YFB1807800???2020YFB1807804, in part by the National Natural Science Foundation of China under Grants 62071067,62001054 and 61771068, in part by the Beijing Municipal Natural Science Foundation under Grant 4182041, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX20200067, and in part by the Ministry of Education and China Mobile Joint Fund MCM20180101, and in part by the Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

1. Guo, L., Yin, H., Wang, Q., Chen, T., Zhou, A., Hung, N.Q.V.: Streaming Session-based Recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Pp.1569–1577. ACM Press, Anchorage, AK, USA (2019). https://doi.org/10.1145/3292500.3330839
2. Chen, T., Yin, H., Chen, H., Yan, R., Nguyen, Q.V.H., Li, X.: AIR: Attentional Intention-Aware Recommender Systems. In: Proceedings of the 35th International Conference on Data Engineering. pp. 304–315. IEEE Press, Macao, China(2019). https://doi.org/10.1109/ICDE.2019.00035

3. Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q.Z., Orgun, M.: Sequential Recommender Systems: Challenges, Progress and Prospects. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 6332–6338. AAAI Press, Macao (2019)
4. Liu, Y., Pham, T.-A.N., Cong, G., Yuan, Q.: An experimental evaluation of point-of-interest recommendation in location-based social networks. Proc. VLDB Endow. **10**, 1010–1021 (2017). https://doi.org/10.14778/3115404.3115407
5. Li, J., Lu, K., Huang, Z., Shen, H.T.: Two Birds One Stone: On both Cold-Start and Long-Tail Recommendation. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 898–906. ACM Press, California (2017)
6. Gao, H., Tang, J., Hu, X., Liu, H.: Content-aware point of interest recommendation on location-based social networks. In: Proceedings of the 29th AAAI conference on artificial intelligence. pp. 1721–1727.AAAI Press, Austin, Texas (2015)
7. Guo, H., Li, H., He, M., Zhao,X.Y., Liu, G.Q., Xu GD.: Joint Factor Model with Content, Social, Location for Heterogeneous Point-of-Interest Recommendation. In: Proceedings of the International Conference on Knowledge Science, Engineering and Management. pp. 613–627. Springer Press, Passau, Germany (2016)
8. Yin, H., Zhou, X., Cui, B., Wang, H., Zheng, K., Nguyen, Q.V.H.: Adapting to user interest drift for POI recommendation. IEEE Trans. Knowl. Data Eng. **28**, 2566–2581 (2016). https://doi.org/10.1109/TKDE.2016.2580511
9. Wang, H., Wang, N., Yeung, D.-Y.: Collaborative deep learning for recommender systems. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1235–1244. ACM Press, Sydney (2015)
10. Kim, D., Park, C., Oh, J., Lee, S., Yu, H.: Convolutional matrix factorization for document context-aware recommendation. In: Proceedings of the 10th ACM conference on recommender systems, pp. 233–240. ACM Press, Boston (2016)
11. Zhang, Y., Yin, H., Huang, Z., Du, X., Yang, G., Lian, D.: Discrete deep learning for fast content-aware recommendation. In: Proceedings of the eleventh ACM international conference on web search and data mining, pp. 717–726. ACM Press, Marina Del Rey (2018)
12. Yin, H., Wang, W., Wang, H., Chen, L., Zhou, X.: Spatial-aware hierarchical collaborative deep learning for POI recommendation. IEEE Trans. Knowl. Data Eng. **29**, 2537–2551 (2017). https://doi.org/10.1109/TKDE.2017.2741484
13. Liu, Q., Wu, S., Wang, L., Tan, T.: Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 194–200. AAAI Press, Phoenix (2016)
14. Liao, J., Liu, T., Liu, M., Wang, J., Wang, Y., Sun, H.: Multi-context integrated deep neural network model for next location prediction. IEEE Access. **6**, 21980–21990 (2018). https://doi.org/10.1109/ACCESS.2018.2827422
15. Cheng, C., Yang, H., King, I., Lyu, M.R.: Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 17–23. AAAI Press, Toronto (2012)
16. Lian, D., Zhao, C., Xie, X., Sun, G., Chen, E., Rui, Y., GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 831–840. ACM Press, New York (2014)
17. Liu, Y., Wei, W., Sun, A., Miao, C.: Exploiting Geographical Neighborhood Characteristics for Location Recommendation. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 739–748. ACM Press, Shanghai (2014)
18. Wang, H., Shen, H., Ouyang, W., Cheng, X.: Exploiting POI-Specific Geographical Influence for Point-of-Interest Recommendation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 3877–3883. IJCAI Press, Stockholm (2018)
19. Liu, B., Xiong, H., Papadimitriou, S., Fu, Y., Yao, Z.: A general geographical probabilistic factor model for point of interest recommendation. IEEE Trans. Knowl. Data Eng. **27**, 1167–1179 (2015). https://doi.org/10.1109/TKDE.2014.2362525
20. Li, X., Jiang, M., Hong, H., Liao, L.: A time-aware personalized point-of-interest recommendation via high-order tensor factorization. ACM Trans. Inf. Syst. **35**, 1–23 (2017). https://doi.org/10.1145/3057283
21. Yao, L., Sheng, Q.Z., Wang, X., Zhang, W.E., Qin, Y.: Collaborative location recommendation by integrating multi-dimensional contextual information. ACM Trans. Internet Technol. **18**, 1–24 (2018). https://doi.org/10.1145/3134438
22. Liu, B., Xiong, H.: Point-of-interest recommendation in location based social networks with topic and location awareness. In: Proceedings of the 2013 SIAM international conference on data mining, pp. 396–404. SIAM, Philadelphia (2013)

23. Li, X., Xu, G.D., Chen, E.H., Li, L.: MARS: A multi-aspect Recommender system for Point-of-Interest. In: Proceedings of the IEEE 31st International Conference on Data Engineering, pp. 1436–1439. IEEE press, South Korea (2015)
24. Wang, H., Fu, Y., Wang, Q., Yin, H., Du, C., Xiong, H.: A location-sentiment-aware recommender system for both home-town and out-of-town users. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1143. ACM Press, Halifax (2017)
25. Xie, M., Yin, H., Wang, H., Xu, F., Chen, W., Wang, S.: Learning Graph-based POI Embedding for Location-based Recommendation. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 15–24. ACM Press, Indianapolis (2016)
26. Zhang, F., Yuan, N.J., Lian, D., Xie, X., Ma, W.-Y.: Collaborative Knowledge Base embedding for recommender systems. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 353–362. ACM Press, San Francisco (2016)
27. Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., Liu, H.: What your images reveal: exploiting visual contents for point-of-interest recommendation. In: Proceedings of the 26th international conference on world wide web, pp. 391–400. ACM Press, Perth (2017)
28. Li, P., Burges, C.J.C., Wu, Q.: Learning to Rank Using Classification and Gradient Boosting. In: Advances in Neural Information Processing Systems 20. p. 10 (2008)
29. Jing He, L.L., Xin Li: Category-aware Next Point-of-Interest Recommendation via Listwise Bayesian Personalized Ranking. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 1837–1843. IJCAI Press, Melbourne (2017)
30. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, pp. 452–461. AUAI Press, Montreal (2009)
31. Rendle, S., Freudenthaler, C.: Improving pairwise learning for item recommendation from implicit feedback. In: Proceedings of the 7th ACM international conference on web search and data mining, pp. 273–282. ACM Press, New York (2014)
32. Feng, S., Li, X., Zeng, Y., Cong, G., Chee, Y.M., Yuan, Q.: Personalized Ranking Metric Embedding for Next New POI Recommendation. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, pp. 2069–2075. AAAI Press, New York (2015)
33. Li, X., Cong, G., Li, X.-L., Pham, T.-A.N., Krishnaswamy, S.: Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 433–442. ACM Press, Santiago (2015)
34. Zhao, S., Zhao, T., Yang, H., Lyu, M.R., King, I.: STELLAR: Spatial-Temporal Latent Ranking for Successive Point-of-Interest Recommendation. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 315–322. AAAI Press, Phoenix (2016)
35. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal Deep Learning. In: Proceedings of the 28th International Conference on Machine Learning. pp. 689–696. ACM Press, Bellevue (2011)
36. Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., Cichocki, A.: EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. IEEE Trans. Cybern. 1–13 (2018). doi:https://doi.org/10.1109/TCYB.2018.2797176
37. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial Cross-Modal Retrieval. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 154–162. ACM Press, Mountain View (2017)
38. Oramas, S., Nieto, O., Sordo, M., Serra, X.: A Deep Multimodal Approach for Cold-start Music Recommendation. ArXiv170609739 Cs. 32–37 (2017). doi:https://doi.org/10.1145/3125486.3125492
39. Zhao, Z., Yang, Q., Lu, H., Weninger, T., Cai, D., He, X., Zhuang, Y.: Social-aware movie recommendation via multimodal network learning. IEEE Trans. Multimed. **20**, 430–440 (2018). https://doi.org/10.1109/TMM.2017.2740022
40. Wang, X., Zhao, Y.-L., Nie, L., Gao, Y., Nie, W., Zha, Z.-J., Chua, T.-S.: Semantic-based location recommendation with multimodal venue semantics. IEEE Trans. Multimed. **17**, 409–419 (2015). https://doi.org/10.1109/TMM.2014.2385473
41. He, R., McAuley, J.: VBPR: visual Bayesian personalized ranking from implicit feedback. In: Proceedings of the 30th AAAI conference on artificial intelligence. pp. 144–150. AAAI Press, Phoenix (2016)
42. Srivastava, N., Mansimov, E., Salakhutdinov, R.: Unsupervised learning of video representations using LSTMs. In: Proceedings of the 32nd international conference on machine learning. pp. 843–852. ACM Press, Lille (2015)
43. Yin, H., Chen, H., Sun, X., Wang, H., Wang, Y., Nguyen, Q.V.H.: SPTF: A Scalable Probabilistic Tensor Factorization Model for Semantic-Aware Behavior Prediction. In: 2017 IEEE International Conference on Data Mining, pp. 585–594. IEEE, New Orleans (2017)

44. Liu, Y., Yang, J.: Improving ranking-based recommendation by social information and negative similarity. Procedia Comput. Sci. **55**, 732–740 (2015). https://doi.org/10.1016/j.procs.2015.07.164
45. Yang, D., Zhang, D., Zheng, V.W., Yu, Z.: Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Trans. Syst. Man Cybern. Syst. **45**, 129–142 (2015). https://doi.org/10.1109/TSMC.2014.2327053
46. Yin, H., Zou, L., Nguyen, Q.V.H., Huang, Z., Zhou, X.: Joint Event-Partner Recommendation in Event-based Social Networks. In: Proceedings of the 34th IEEE International Conference on Data Engineering. pp.929-940. IEEE Press, Paris, France (2018)
47. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448–456. ACM Press, San Diego (2011)
48. Mnih, A., Salakhutdinov, R.R.: Probabilistic matrix factorization. In: Proceedings of the 20th international conference on neural information processing systems. pp. 1257–1264. MIT Press, Vancouver (2007)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.