Spatio-Temporal Attention based Recurrent Neural Network for Next Location Prediction

Basmah Altaf
CEMSE
KAUST
City, Country
Thuwal, Saudi Arabia
basmah.altaf@kaust.edu.sa

Lu Yu
CEMSE
KAUST
Thuwal, Saudi Arabia
lu.yu@kaust.edu.sa

Xiangliang Zhang
CEMSE
KAUST
Thuwal, Saudi Arabia
xiangliang.zhang@kaust.edu.sa

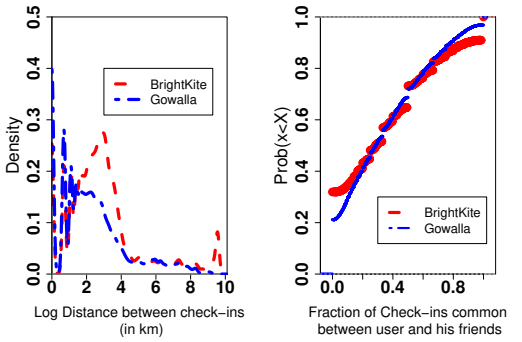
***Abstract*—With the advances in technology and smart devices, more and more attention has been paid to model spatial correlations, temporal dynamics, and friendship influence over point-of-interest (POI) checkins. Besides directly capturing general user's checkin behavior, existing works mostly highlight the intrinsic feature of POIs, *i.e.,* spatial and temporal dependency. Among them, the family of methods based on *Markov chain* can capture the instance-level interaction between a pair of POI checkins, while *recurrent neural network* (RNN) based approaches (state-of-the-art) can deal with flexible length of checkin sequence. However, the former is not good at capturing high-order POI transition dependency, and the latter cannot distinguish the exact contribution of each POI in a historical checkin sequence. Moreover, in recurrent neural networks, local and global information is propagated along the sequence through one bottleneck i.e., hidden states only.**

In this work, we design a novel model to enforce contextual constraints on sequential data by designing a spatial and temporal attention mechanisms over recurrent neural network that leverages the importance of POIs visited by users in given time interval and geographical distance in successive checkins. Attention mechanism helps us to learn which POIs bounded by time difference and spatial distance in user checkin history are important for the prediction of next POI. Moreover, we also consider periodicity and friendship influence in our model design. Experimental results on two real location based social networks Gowalla, and BrightKite show that our proposed method outperforms the existing state-of-the-art deep neural network methods for next POI prediction and understanding user transition behavior. We also analyze the sensitivity of parameters including context window for capturing sequential effect, temporal context window for estimating temporal attention and spatial context window for estimating spatial attention respectively.

***Index Terms*—spatiotemporal; attention; deep learning; sequence modeling; memory networks**

I. Introduction

With the availability of smart devices, users tend to share their point of interests (POI) in social networks via check-ins. The history of users check-in behavior is rich in temporal and spatial contexts, and personalized recommendation systems are built using sequential check-in history for helping user to navigate to POIs at next time point. Modeling user transition preferences finds its application not only for personalized POI prediction but also in adequate resource planning, budgeting, providing better services, and transportation planning.



(a) Density of distance in consecutive checkins

(b) CDF of fraction of common checkins between user and his friends

Fig. 1. Influence of distance and friendship on user check-in behavior

POI recommendation is influenced by multiple aspects, such as spatial influence, temporal influence, friends (social) influence and time of check-in [1]. Usually, users prefer to checkin at close POIs. For example, when facing two grocery stores with different distances, users would tend to visit the closer one from his current location $d1 < d2$. This adds spatial constraint in modeling check-in sequences. Similarly, check-ins made in close time-interval may greatly influence the next POI checkin, e.g., going for coffee right after having lunch, and going to bar after dinner, etc. Moreover, the social aspect may influence the user $u1$ to visit the places that were already visited/recommended by his friend ($u2$). Figure 1(a) illustrates the inverse relation of the distance to the check-in frequency, thus showing that users prefer to checkin at nearby places. It was obtained by plotting the density of distance between the chronologically sorted consecutive check-ins of each user. To illustrate how user's friends influence his behavior of POI checkins, we estimate for each user the fraction of checkins of POIs commonly visited by him and his friends at unique places, and plot cumulative distribution in Figure 1(b). We see a steep slope in empirical cumulative distribution function curve showing the evidence that friends have an imminent influence on user checkin behavior.

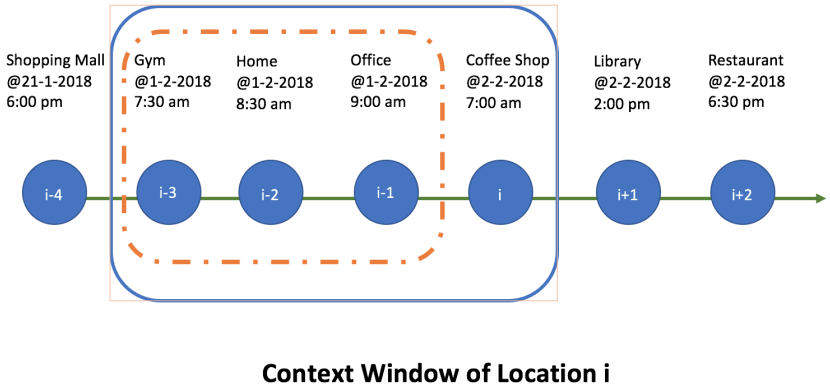


Fig. 2. An example of the temporal context window of a user's i^{th} visited location. The window includes 3 locations visited before the i^{th} location due to temporal difference w.r.t. current checkins is within w hours. Here $w = 24$ hours, for example.

With the advances in deep learning, recurrent neural networks (RNN) are considered as state of the art for modeling sequential data. It assumes that the data in a sequence are equally spaced in time. However, in check-in sequences, time difference between check-ins is not constant (see Figure 2 as an example). We observe that users do not make check-ins at regular time difference w.r.t previous check-in. Spatio-Temporal Recurrent Neural Network (ST-RNN) [9], Semantics Enriched Recurrent Model (SERM) [13], and DeepMove [4] have been proposed for next POI prediction by adapting recurrent neural networks in their approach. Nonetheless, all these techniques suffer from long range dependency problem in capturing the importance of far previous inputs to predict later in the sequence.

Based on the above observations, we propose to take both **geographical** and **temporal dependency** into account, and develop a novel ranking model based on attention over recurrent neural network. Specifically, we instantiate a *spatial attention* layer to capture the geographical correlation between the past and future POIs based on their distance. Meanwhile, temporal dynamics in existing POI sequence is tend to be modeled by a *temporal attention* layer over the correlation and significance of POIs checked in short duration, while the recurrent neural network captures the contextual dependency. We also model user **friendship** and **POI co-occurence** using word2vec pre-trained embeddings for users and POIs on friendship edges and POI sequences respectively. Also, to model **periodicity**, we consider time of visit for each POI in our model. Finally, we jointly optimize the next POI prediction task using back-propagation through time (BPTT) and bayesian pairwise ranking (BPR) as the ranking algorithm.

II. Related Work

Ranking techniques have been deployed extensively in recommendation systems for predicting user's preference for books, items, locations etc. Rende et al. [11] proposed a bayesian pairwise ranking (BPR) approach with relative preference of implicit items over non-observed items for each user. However, the main bottleneck in ranking techniques is that all users act independent to each other, and the ordering of each pair of items for a specific user is independent of the ordering of every other pair.

Many POI recommendation models are built on top of matrix factorization by focusing on different aspects such as geographical influence [14], temporal influence [5], and semantic influence [6]. However, all of these ignore sequential dependency. Tensor Factorization (TF) has been successfully applied for time-aware recommendation as well as modeling spatial and temporal information [15]. In TF, both time bins and locations are regarded as additional dimensions in the factorized tensor, which leads to the cold start problem in behavior prediction with new time bins.

It is however, important to consider the successive checkins history since there is correlation found in successive checkins [2]. Different techniques have been proposed for sequential modeling including Markov chain model incorporation in collaborative filtering, recurrent neural network, and word2vec framework. However, all Markov chains based methods assume strong independence among different components and only consider the last POI when modeling POI sequences. In contrast, the user's choice for current POI may be influenced by more than one POIs visited in the past.

Recently, recurrent neural networks have gained remarkable attention and are state of the art in modeling sequence history and transition of user's movement. To predict next location of a user at a specific time t based on his checkin history, ST-RNN models time difference and spatial difference between consecutive checkins [9] by partitioning the space and time into bins, and learns a transition matrix for every temporal and spatial bin. However, it considers only a fixed number of previous inputs, thus ignoring long range dependencies. Moreover, we need to manually set the bounds of transition matrices, thus impractical for different datasets. SERM [13] learns POI dynamics by embedding different locations, users, time-bins and text. However, this work does not study spatial or temporal influence on user check-ins, and ignores long-range dependencies by splitting one user checkin sequence into multiple, and treat them independently. To learn periodic

dependency of user check-in behavior, DeepMove [4] learns periodic behavior dependency from user past checkins for predicting next POI. Our work on the other hand learns spatial difference and time interval dependency using spatio-temporal attention over recurrent neural network for modeling user check-in sequence. Moreover, we also consider friendship influence and periodicity in user checkin behavior.

III. Preliminaries

Let $u \in U$ be a set of N unique users such that $U = \{u_1, u_2, ..., u_N\}$, and let L be a set of M unique point of interests (POIs) such that $L = \{l_1, l_2, ..., l_M\}$ in a location based social network (LBSN).

Definition 1 (User Social Links) Let $G = (V, E)$ be a user social links graph, where each node V represents a unique user u and edges E show the social ties.

Definition 2 (Sequence) Let a sequence q_u be a set of POIs visited by user u in chronological order $q = \{l_1, l_2, l_3, ...l_K\}$ such that the time of visit for each POI is $t_1 < t_2 < \cdots < t_K$ where K is variable length of each sequence. We define a **context window** w that determines how many POIs to be included as context POIs for the given POI. For each user, we define $Q_u =< q_u^i >_{i=1}^{|Q_u|}$ is a set of sequences collection for each user, and all user sequences can be represented by $Q =< Q_u >_{u=1}^{N}$, where N is count of unique users.

Definition 3 (Time Stamp Sequence) Let a sequence t_u be a set of timestamps at which user visited POIs in chronological order $t_u = \{t_1, t_2, \ldots, t_K\}$. For each user, we define $T_u =< t_u^i >_{i=1}^{|Q_u|}$, a set of time-stamps collection of sequence of visited POIs.

Definition 4 (Spatial Sequence) Let a sequence s_u be a set of latitude and longitude points that represent geographical coordinates of POIs visited $s_u = \{(lat_1, long_1), \ldots, (lat_K, long_K)\}$. For each user, we define $S_u =< s_u^i >_{i=1}^{|Q_u|}$, a set of geographical coordinates of sequence of visited POIs.

Definition 5 (Temporal Context Window) Let τ_t be a context window that defines the number of hours, in which checkins are made before the current POI to be considered as **temporal context** for next POI prediction.

Definition 6 (Spatial Context Window) Let τ_l be a spatial context window that defines geographical distance in km from the current POI to checked-in POIs in the past to qualify those POIs to be considered as spatial context for next POI prediction.

Definition 7 (Time bin) Let l be the time bin (hour of visit) for each poi. We represent time of visit in 48 bins, $0 - 23$ for weekdays and $24 - 47$ for weekends.

Problem Definition: Given user social graph G, user checked-in POIs sequence Q, time of check-in for each POI sequence T, geographical coordinates of each POI sequence S, context window w, temporal context window τ_t, spatial context window τ_l, and time bin of visit l, we learn compact representation of users, time bins and POIs in latent space to predict next POI for each user.

IV. Proposed Model

A. Embedding

This layer maps object ids to vector of real numbers and define semantic similarity of entities (users, time bins and POIs) in vector space. We initialize the user and POI embedding layer with pre-trained embeddings using word2vec [10] proposed by Mikolov learnt on users social links graph, and user POI sequences respectively. In our model, we use four different embedding layers, one for user id, one for target POI id, one for context embedding of sequence of POIs and for time-bins of sequence of POIs. We then concatenate the representation of time-bins and POIs of checkin sequence, to learn compact representation of user checkin.

B. Recurrent Neural Networks

In our experiments we use Gated Recurrent Units (GRUs) which are a more robust variant of recurrent neural networks (RNN) and work better in capturing long term dependencies. A GRU has two gates, a reset gate (r) to determine how to combine the current input x_t and previous memory h_{t-1}, while an update gate (z) to define how much of the previous information to keep.

C. Attention Mechanism

RNN compresses all the visited POIs in check-in sequence into a fixed length vector (last hidden state) to determine the next POI at each time step. Attention mechanism, on the other hand, adepts to understand the sequence of POI check-ins in order to select only the important and relevant POIs for next POI prediction at each time step. It takes all the previous hidden states as input to compute the probability distribution of previous checked-in POIs and generates a context vector as a weighted sum of visited POIs for next POI prediction. By using this approach, it is possible for model to capture somewhat global information rather than to infer based on one hidden state only, representing the whole sequence.

The context vector is built by following the steps given in eq. (1)-(4).

Step 1: Compute the similarity score of each previous hidden state with the current hidden state

$$e_{ij} = a(h_i, h_j) = W_s h_i + W_h h_j \tag{1}$$

where $j = \{1, ..., i-1\}$, i is the current time-step and a is a feed-forward neural network, trained jointly with other components of our proposed model.

Step 2: Using softmax function, the alignment scores are normalized.

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{i-1} exp(e_{ik})} \tag{2}$$

Step 3: The context vector is a weighted sum of hidden states (h_j) and normalized alignment scores (α_{ij}).

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij} h_j \tag{3}$$

Step 4: The context vector is then concatenated with the target hidden state.

$$a_i = f(c_i, h_i) = tanh(W_c[c_i; h_i]) \tag{4}$$

D. Spatio-Temporal Attention over Gated Recuurent Unit (STA-GRU)

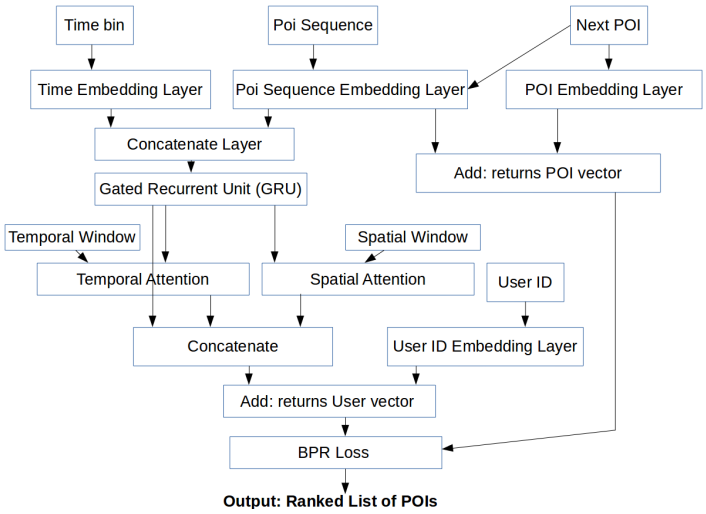


Fig. 3. Architecture of Spatio-Temporal Attention based GRU (STA-GRU) for next POI prediction.

In our model as shown in Fig 3, we use two modes of attention mechanism namely *Temporal Attention* and *Spatial Attention*, respectively. In both modes, the attention component takes $t-1$ hidden states of GRU as arguments $\{h_1, h_2, \ldots, h_{t-1}\}$ and a context vector h_t and context window τ in our case. It returns a vector z, which is the summary of all the previous hidden states focusing on the information linked to context h_t and context window τ. More formally, it returns a weighted arithmetic mean of the hidden states of GRU $\{h_i\}_{i=1}^{t-1}$ and the weights are chosen according to the relevance of each h_i given the context h_t.

Temporal Attention. Subsequent checkins in close time interval are highly correlated [15] and the correlation decreases as the interval between two checkins is increased. In temporal attention, we learn weights using attention mechanism for each of the visited POIs in past in a temporal context window τ_t using (5).

$$h_j \in H \quad where\ \Delta_t = |(t_j - t_h)| \leq \tau_t \tag{5}$$

Spatial Attention. Places that are nearby have high tendency to get checked in by user, rather than far-away places [15].

With spatial attention mechanism, we enforce the spatial constraints to learn contributions of nearby POIs visited by user in the past using (6) to determine next POI.

$$h_j \in H \quad where\, \Delta_s = |(l_j - l_h)| \leq \tau_l \tag{6}$$

such that $l = (latitude, longitude)$ and **haversine** distance is used to calculate distance between two points assuming that the earth is spherical.

After getting the temporal attention context vector c_t and spatial attention context vector c_s, we concatenate the last hidden state of GRU that captures user's dynamic pattern of checkins with these two vectors and add it to user id embedding vector to obtain final representation for user u as shown in (7).

$$\vec{p} = [\vec{h_t}; \vec{c_t}; \vec{c_s}] + \vec{u} \tag{7}$$

Similarly, we add the POI embeddings obtained from two separate POI embedding layers to get final representation of each POI as shown in (8).

$$\vec{q} = \vec{q_i} + \vec{q_{c_i}} \tag{8}$$

Finally, the prediction of STA-GRU can be inferred by calculating inner product of user p and POI q representations as given in (9).

$$o_{u,t,q} = \mathbf{p} \cdot \mathbf{q} \tag{9}$$

E. Parameter Learning

Given a collection of user-POI check-ins Q, the objective is to represent users, time-bins and POIs in a new unified embedding space, where the new space captures the latent mobility patterns in the raw data. We jointly learn the embeddings for users and POIs using a bayesian pairwise ranking (BPR) loss and back propagation through time (BPTT) and formulate the training objective function as given in (10)

$$J_t = \sum_{Q} (1 - \sigma(o_{u,t,q} - o_{u,t,q'})) + \lambda ||(\theta)||^2 \tag{10}$$

where q' is a negative location sample, λ is the regularization parameter and θ indicates all the parameters.

V. Experiments and Analysis

A. Datasets

We conduct experiments on two location based online social networks namely Gowalla (GW) [8] and BrightKite (BK) [7]. In both datasets, each checkin comprises of userid, time, latitude, longitude, and location id. We filter the POIs checked-in by less than 15 users, and users who have visited less than 10 POIs. After pre-processing, statistics of both the datasets are given in Table I.

TABLE I
Dataset Statistics

Dataset	Gowalla (GW)	BrightKite (BK)
Duration	Feb 2009- Oct 2010	Apr 2008 - Oct 2010
#Users $\lvert\mathcal{V}\rvert$	37787	6793
#Locations $\lvert\mathcal{P}\rvert$	55803	11932
#Check-insv$\sum_{c,v} X$	2123750	595693
#Friendship edges $\lvert A\rvert$	1900654	428156

In both datasets, we keep the last POI visited by user in the test set, while the former part of user POI checkin sequence in the training. For evaluation, we use popular ranking metrics

TABLE II
PERFORMANCE COMPARISON ON GOWALLA AND BRIGHTKITE DATASETS

Dataset	Method	recall@1	recall@50	recall@100	ndcg@5	ndcg@50	ndcg@100	AUC	MRR
Gowalla	FPMC	0.1067	0.312	0.3581	0.1549	0.1838	0.1912	0.9418	0.1509
(**GW**)	BPR	0.0758	0.4760	0.5539	0.1550	0.2173	0.2299	0.9805	0.1512
	GRU	0.0927	0.4928	0.5644	0.1901	0.2259	0.2450	0.9813	0.2231
	ST-RNN (2016)	0.0914	0.4722	0.5579	0.1599	0.2198	0.2359	0.9540	0.2111
	SERM (2017)	0.0924	0.4838	0.5621	0.1732	0.2209	0.2408	0.9602	0.2153
	DeepMove (2018)	0.1132	0.5010	0.5679	0.1932	0.2336	0.2493	0.9818	0.2391
	STA-GRU	**0.1295**	**0.5609**	**0.6231**	**0.2215**	**0.2597**	**0.2768**	**0.9852**	**0.2612**
	% **Improvement**	14.39%	11.95%	9.72%	14.64%	11.17%	11.03%	0.34%	9.24%
BrightKite	FPMC	0.1569	0.666	0.7549	0.2515	0.3329	0.3473	0.9711	0.2475
(**BK**)	BPR	0.2380	0.7742	0.8192	0.4216	0.4707	0.4780	0.9780	0.3839
	GRU	0.2532	0.7950	0.8365	0.4691	0.4833	0.4971	0.9782	0.4325
	ST-RNN (2016)	0.2431	0.7795	0.8253	0.4522	0.4745	0.4798	0.9512	0.4006
	SERM (2017)	0.2698	0.7998	0.8355	0.4629	0.4891	0.4945	0.9635	0.4310
	DeepMove (2018)	0.2927	0.8042	0.8500	0.4808	0.5092	0.5189	0.9794	0.4640
	STA-GRU	**0.3229**	**0.8561**	**0.8972**	**0.5269**	**0.5413**	**0.5514**	**0.9832**	**0.4893**
	% Improvement	10.31%	6.45%	5.55%	9.58%	6.3%	6.26%	0.38%	5.45%

namely **recall@k**, **ndcg@k**, **area under the curve (AUC)** and **mean reciprocal rank (MRR)**.

B. Comparison Methods

We compare our proposed model with the following:

Factorizing Personalized Markov Chains (FPMC) [12]. introduces a personalized transition matrix and factorizes the transition cube using tensor decomposition.

Bayesian Pairwise Ranking (BPR) [11]. an implicit feedback method by using a pairwise item preferences

Gated Recurrent Unit (GRU) [3]. a robust variant of RNN and has an in-built memory along with gates to determine what information should be passed to the output.

Spatio Temporal Recurrent Network (ST-RNN) [9]. models the user sequence of checkins by considering the spatial and temporal contexts using transition matrices of spatial and time differences between successive checkins.

Semantics Enriched Recurrent Model (SERM*) [13]. a variant of SERM, which only models location, time, and user factors without using textual information.

Attentional Recurrent Networks (DeepMove) [4]. learns the periodic contribution of user checkin history by attention mechanism, on sequence-level.

C. Analysis of Experimental Results

Table II shows performance of our proposed method STA-GRU on two datasets. We observe that BPR outperforms FPMC on our data sets, while GRU is better than BPR, as it considers sequence history. Deep learning based methods ST-RNN and SERM do not show good performance as compared to GRU because for each prediction, they consider only fixed context from the past, and thus do not capture long-term dependencies. DeepMove on the other hand is able to capture long term periodic dependencies using attention mechanism on the past sequences representation, and thus performs better than ST-RNN and SERM. Our proposed method STA-GRU outperforms all these methods because it captures long term dependency using GRU, and applies attention in the temporal and spatial context to give more weights to POIs that were checked in closer in time, and are located nearby, along with modeling the friends influence and periodicity.

D. Influence of the Attention Mechanism

TABLE III
COMPARISON OF DIFFERENT ATTENTION MECHANISMS

Dataset	Method	recall@100	ndcg@100	AUC	MRR
GW	GRU	0.5644	0.2450	0.9813	0.2231
	GRU+SA	0.6073	0.2531	0.9822	0.2562
	GRU+TA	0.6158	0.2694	0.9841	0.2597
	GRU + STA	**0.6231**	**0.2768**	**0.9852**	**0.2612**
BK	GRU	0.8365	0.4971	0.9782	0.4325
	GRU+SA	0.8623	0.5328	0.9809	0.4731
	GRU+TA	0.8759	0.5473	0.9825	0.4802
	GRU + STA	**0.8972**	**0.5514**	**0.9832**	**0.4893**

In table III, we show the results of models based on GRU network with different attention mechanisms. GRU+SA considers only spatial attention while GRU+TA considers only temporal attention. GRU+STA combines both spatial and temporal attentions. From the results, we can observe that GRU+STA outperforms the other three methods, which proves the effectiveness of the spatio-temporal attention mechanism. Besides, the temporal attention shows better performance than the spatial-level attention and basic GRU.

E. Finding and Analysis of Optimal Context Width

TABLE IV
ANALYSIS OF CONTEXT WINDOW LENGTH ON GOWALLA AND BRIGHTKITE DATASETS

Dataset	length	recall@100	ndcg@100	AUC	MRR
GW	2	0.6204	0.2595	0.9835	0.2606
	5	**0.6231**	**0.2768**	**0.9852**	**0.2612**
	8	0.6075	0.2506	0.9827	0.2547
BK	2	0.8629	0.5391	0.9851	0.4744
	4	**0.8972**	**0.5514**	**0.9832**	**0.4893**
	6	0.8579	0.5292	0.9824	0.4697

TABLE V
ANALYSIS OF TIME WINDOW ON GOWALLA AND BRIGHTKITE DATASETS

Dataset	time window	recall@100	ndcg@100	AUC	MRR
GW	12	0.5731	0.2518	0.9820	0.2432
	24	0.5877	0.2572	0.9827	0.2586
	48	**0.6231**	**0.2768**	**0.9852**	**0.2612**
	60	0.5743	0.2536	0.9823	0.2579
BK	12	0.8512	0.5326	0.9821	0.4325
	24	**0.8972**	**0.5514**	**0.9832**	**0.4893**
	48	0.8743	0.5379	0.9825	0.4782
	60	0.8023	0.4933	0.9815	0.4249

With the increase of context length until the optimal length 5 for GW and 4 for BK, we get increase in performance metrics; however, after the optimal length, there is sudden performance degradation with the increase in context length as shown in Table IV. This indicates that performance gain happens with the increase in context window until adding more context brings noise.

F. Finding and Analysis of Optimal Temporal Width

To select best value for temporal context width, we fix all other parameters, and vary temporal context window, and choose the one that gives best performance as shown in Table V. On GW dataset, we get better results when using temporal window of 48 hours, while for BK, we achieve better performance when using temporal context window of 24 hours respectively.

G. Finding and Analysis of Optimal Spatial Width

We show the results of fine-tuning on our datasets for optimal spatial window to apply spatial attention in Table VI. To select best value for spatial context width, we fix all other parameters, and vary spatial context window from low to high values until the increase in distance does not contribute to co-occurrence of POIs and adds noise.

TABLE VI
ANALYSIS OF SPATIAL WINDOW ON GOWALLA AND BRIGHTKITE DATASETS

Dataset	spatial window	recall@100	ndcg@100	AUC	MRR
GW	3	0.6053	0.2603	0.9842	0.2548
	6	0.6106	0.2715	0.9848	0.2602
	7	**0.6231**	**0.2768**	**0.9852**	**0.2612**
	10	0.5738	0.2516	0.9825	0.2102
BK	3	0.8769	0.5478	0.9827	0.4759
	6	**0.8972**	**0.5514**	**0.9832**	**0.4893**
	7	0.8751	0.5456	0.9821	0.4753
	10	0.8598	0.5237	0.9803	0.4691

VI. CONCLUSION

We introduce the spatio-temporal attention over gated recurrent units (STA-GRU) for POI sequence modeling; a novel neural network architecture based on a self-attention mechanism that we believe to be particularly well-suited for modeling user transition behavior and next POI prediction. Through experiments, we show that STA-GRU outperforms existing state of the art models for POI sequence modeling on Gowalla and BrightKite benchmarks. Long range dependencies are still tricky in recurrent neural networks even with the gating mechanism. Therefore, we introduce spatial and temporal attention over recurrent neural networks.

ACKNOWLEDGMENT

This work is supported by King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

REFERENCES

[1] Ramesh Baral, Dingding Wang, Tao Li, and Shu-Ching Chen. Geotecs: exploiting geographical, temporal, categorical and social aspects for personalized poi recommendation. In *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on*, pages 94–101. IEEE, 2016.

[2] Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, volume 13, pages 2605–2611, 2013.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1459–1468. International World Wide Web Conferences Steering Committee, 2018.

[5] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 93–100. ACM, 2013.

[6] Bo Hu and Martin Ester. Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 25–32. ACM, 2013.

[7] Jure Leskovec and Andrej Krevl. Snap datasets: Brightkite checkin data set. http://snap.stanford.edu/data, June 2011.

[8] Jure Leskovec and Andrej Krevl. Snap datasets: Gowalla checkin data set. http://snap.stanford.edu/data, June 2011.

[9] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 194–200, 2016.

[10] 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.

[11] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.

[12] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820. ACM, 2010.

[13] Di Yao, Chao Zhang, Jianhui Huang, and Jingping Bi. Serm: A recurrent model for next location prediction in semantic trajectories. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2411–2414. ACM, 2017.

[14] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 325–334. ACM, 2011.

[15] Shenglin Zhao, Tong Zhao, Haiqin Yang, Michael R Lyu, and Irwin King. Stellar: Spatial-temporal latent ranking for successive point-of-interest recommendation. In *AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 315–322, 2016.