The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Spatiotemporal Activity Modeling Under Data Scarcity: A Graph-Regularized Cross-Modal Embedding Approach

Chao Zhang,[1] **Mengxiong Liu,**[1] **Zhengchao Liu,**[1] **Carl Yang,**[1] **Luming Zhang,**[2] **Jiawei Han**[1]
[1]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2]EmoKit Tech Co., Ltd., Beijing, China
[1]{czhang82, mliu60, zliu80, jiyang3, hanj}@illinois.edu} [2]zglumg@gmail.com

Abstract

Spatiotemporal activity modeling, which aims at modeling users' activities at different locations and time from user behavioral data, is an important task for applications like urban planning and mobile advertising. State-of-the-art methods for this task use cross-modal embedding to map the units from different modalities (location, time, text) into the same latent space. However, the success of such methods relies on data sufficiency, and may not learn quality embeddings when user behavioral data is scarce. To address this problem, we propose BRANCHNET, a spatiotemporal activity model that transfers knowledge from external sources for alleviating data scarcity. BRANCHNET adopts a graph-regularized cross-modal embedding framework. At the core of it is a main embedding space, which is shared by the main task of reconstructing user behaviors and the auxiliary graph embedding tasks for external sources, thus allowing external knowledge to guide the cross-modal embedding process. In addition to the main embedding space, the auxiliary tasks also have branched task-specific embedding spaces. The branched embeddings capture the discrepancies between the main task and the auxiliary ones, and free the main embeddings from encoding information for all the tasks. We have empirically evaluated the performance of BRANCHNET, and found that it is capable of effectively transferring knowledge from external sources to learn better spatiotemporal activity models and outperforming strong baseline methods.

Introduction

Spatiotemporal activity modeling aims at modeling people's activities at different geographical locations and temporal points. It is an important task for a wide variety of real-life applications. For example: 1) what are the typical leisure activities around 8pm on the 5th Avenue? 2) which areas in the New York City do people usually visit for shopping electronic products? Answering such questions is highly useful for applications ranging from mobile advertising to urban planning and tourism recommendation. Recent years have witnessed inspiring results of leveraging user behavioral data for this problem (Sizov 2010; Kling et al. 2014; Yin et al. 2011; Zhang et al. 2017b; 2016; 2017a; Abdelhaq, Sengstock, and Gertz 2013; Feng et al. 2015). With the ubiquitous access to the mobile Internet, people are increasingly sharing their activities in the physical world on social media platforms (*e.g.*, Twitter, Facebook, Instagram). Every day, billions of people go to different places (restaurants, malls, airports, *etc.*) in the world and leave behind them massive trace data on social media platforms (Cheng et al. 2011; Cranshaw et al. 2012; Zhang et al. 2017b). Complementary to conventional sensing data (Yao et al. 2017), such socially sensed behavioral data consist of not only rich location and time information but also the textual descriptions of people's activities, thus serving a multi-dimensional *what-where-when* data source for understanding people's activities in the physical world.

State-of-the-art spatiotemporal activity models (Xie et al. 2016; Zhang et al. 2017b; 2017c) rely on cross-modal embedding. They embed the units from different modalities (location, time, and keywords) into the same latent space to derive their vectorized representations; and two units that are highly correlated (*e.g.*, the 5th Avenue location and the keyword 'shopping') tend to have close embeddings. Once the embeddings are learned, the typical activities at different locations and time can be easily retrieved based on vector similarities. Compared with earlier topic model-based methods (Sizov 2010; Kling et al. 2014; Yin et al. 2011; Hong et al. 2012; Yuan et al. 2013), such a cross-modal embedding approach does not impose assumptions about the distributions of different modalities, and demonstrate excellent scalability (Zhang et al. 2017b).

Unfortunately, spatiotemporal activity modeling based on cross-modal embedding largely relies on data sufficiency. They often require a sufficient amount of user behavioral data, such that the learned embeddings can well capture the correlation structures between location, time, and text, and thus generalize well for prediction tasks. Many practical scenarios, however, involve a limited amount of user behavioral data. For instance, while massive social media records are published in large cities like New York City, there may be a small amount of them in a less populous town, yet it is still important to build spatiotemporal activity models for such towns. As another example, to understand people's activities in a specific time period (*e.g.*, Christmas holiday), it is desirable to put emphasis on the records during that period, but again, one has to deal with the data scarcity problem as the training data for that specific period could be small.

We study the problem of modeling people's spatiotempo-

ral activities under data scarcity. While the given user behavioral data is limited, we claim that there are external sources that reflect the correlations between different units and can be useful for the spatiotemporal activity modeling task at hand. For example, given a small corpus that consists of merely thousands of user behavioral records, two keywords 'shop' and 'store' may not occur frequently enough to push their embeddings close to each other. However, their correlations may be reflected in external sources like WordNet (Miller 1995), which provides evidence to conclude the two keywords are semantically correlated and should have close embeddings.

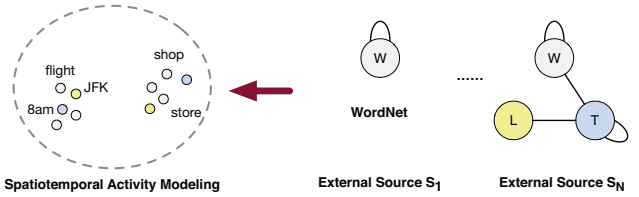


Figure 1: Transferring knowledge from external sources $S_1, \dots S_N$ for spatiotemporal activity modeling.

To transfer the knowledge from external sources, we propose a novel method named BRANCHNET. For each external source, we first use a heterogeneous graph to encode the knowledge between different units in that source, such that two correlated units are connected with an edge, and the edge weight represents the correlation strength. With multiple correlation graphs from different external sources, we design a graph-regularized cross-modal embedding approach that maps location, time, and text into the latent space. Specifically, we learn the cross-modal embeddings in a multi-task framework: 1) one main task of reconstructing user behaviors for the given main corpus; and 2) multiple auxiliary tasks of preserving the correlation graph structures for the external sources. To effectively leverage external knowledge, we design a main embedding space that is shared by the main task and all the auxiliary tasks. The main embedding space bridges user behavior reconstruction with correlation graph embedding, such that the external knowledge can be leveraged to guide cross-modal embedding and alleviate data scarcity. In the mean time, for each specific task, we design a task-specific branched embedding space, which learns task-specific embeddings of different units. Such branched embeddings capture the discrepancies between the main task and the auxiliary ones and free the main embeddings from encoding information for all the tasks.

Our contributions are summarized as follows:

- We study the problem of modeling people's activities from a limited amount of user behavior data, with the help of external sources. To the best of our knowledge, this is the first work that attempts to address the data scarcity problem for spatiotemporal activity modeling.
- We design a novel graph-regularized cross-modal embedding model, which uses a main embedding space to reflect the semantics shared by different tasks, as well as multiple task-specific embeddings to capture the discrepancies of different tasks.
- We have conducted extensive experiments on a number of real-life datasets. Our experimental results show that, compared with state-of-the-art methods, BRANCHNET can better transfer knowledge from relevant external sources and achieves better performance for spatiotemporal activity predictions.

Problem Definition

For the spatiotemporal activity modeling problem, we consider a corpus $\mathcal{C}$ of user behavioral data (*e.g.*, geo-tagged tweets). Each record $r \in \mathcal{C}$ contains a location, a timestamp, and a text message, thus reflecting a user's activity at a specific location and timestamp. Formally, each user behavior record is defined as follows.

Definition 1 (User Behavior Record) *A user behavior record r is described by a tuple $\langle t_r, l_r, m_r \rangle$ where: (1) l_r is a two-dimensional vector that represents the user's location when r is created; (2) t_r is the timestamp when t_r is created; and (3) m_r is a bag of keywords denoting the text message of r.*

To address space and time continuity, we partition the space into equal-sized grids and map a raw GPS location l_r into one of those grids; similarly, for the raw timestamp t_r, we map t_r to some hour in a day and obtain 24 different possible values accordingly. With the corpus $\mathcal{C}$, the task of spatiotemporal activity modeling aims to model people's spatiotemporal activities. In practical applications, however, the given user behavior corpus $\mathcal{C}$ may be small, making it almost impossible to generate reliable spatiotemporal activity models from $\mathcal{C}$ alone. To address such data scarcity, we assume there are N external sources $\mathcal{S}_1, \dots, \mathcal{S}_N$. Each source S_n specifies the correlations between different units with a graph, defined as follows:

Definition 2 (Correlation Graph) *A correlation graph G is an undirected graph $G = (V, E)$ where V is the node set and E is the edge set. Each node $v \in V$ corresponds to a unit (location, time, or keyword) in the corpus $\mathcal{C}$; and each edge $e = (v_i, v_j)$ means two units v_i and v_j is correlated, and the edge weight w_e specifies the correlation strength.*

Note that a correlation graph G may specify the correlations for only a subset of the units in the corpus $\mathcal{C}$. For instance, suppose we use the WordNet (Miller 1995) corpus as an external knowledge source, the corresponding graph G may only involve the keywords that have appeared in the WordNet corpus and their correlations.

With the corpus $\mathcal{C}$ and the external sources $\mathcal{S}_1, \dots, \mathcal{S}_N$, the spatiotemporal activity modeling task aims at modeling people's activities at different locations and time. Given any two of the three factors (location, time, and text), the result spatiotemporal activity model is expected to predict the remaining one, *e.g.*: (1) What are the typical activities at a specific location and time? (2) Given an activity and time, where does this activity occur? and (3) Given an activity and a location, when does the activity occur?

The BRANCHNET Model

In this section, we introduce our proposed BRANCHNET model. We first give an overview of BRANCHNET, and then describe the details of the main task of reconstructing user behaviors and the auxiliary tasks of embedding correlation graphs. Finally, we present the optimization procedure for learning the parameters in BRANCHNET.

Model Overview

To model people's spatiotemporal activities, the key idea of BRANCHNET is to map the units from different modalities into the same latent space, such that correlated units have close embeddings. For example, if the keyword 'shop' occurs frequently around the 5th Avenue area, the embeddings of 'shop' and the 5th Avenue area are encouraged to be close to each other. Given the corpus $\mathcal{C}$ of people's behavioral data, we learn the embeddings of all spatial, temporal, and textual units such that users' behaviors in $\mathcal{C}$ can be reconstructed as much as possible. We call the task of reconstructing user behavioral data in $\mathcal{C}$ the *main task*.

As aforementioned, since the corpus $\mathcal{C}$ may be too small to learn quality cross-modal embeddings, we introduce *auxiliary tasks* for different external sources $S_1, \ldots, S_N$. Each external source S_n provides a correlation graph that specifies the correlations among different units. Hence, for each auxiliary task, we learn the embeddings of different units such that their correlations in the corresponding graph can be preserved.

Figure 2 shows the overall architecture of BRANCHNET. As shown, the units in the given corpus $\mathcal{C}$ have two versions of embeddings: the *main* embeddings and the *auxiliary* embeddings. For the main task of reconstructing user behaviors in $\mathcal{C}$, we directly use the main embeddings for activity prediction (described in detail soon). Meanwhile, we concatenate the main embedding and the auxiliary embedding together for each auxiliary task. As each auxiliary task has an task-specific embedding space to capture the characteristics of that task, the main embeddings have more flexibility to capture the relationships of the units that are shared across all the tasks.

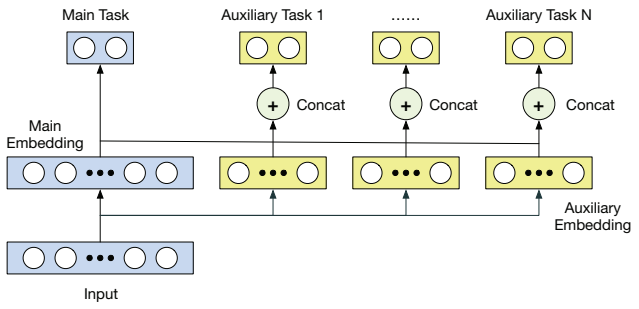


Figure 2: The BRANCHNET model: the main embedding space is shared by the main task and the auxiliary tasks; while each auxiliary task also has a task-specific embedding space to capture the unique structure for that task.

Main Task: Reconstructing Behavior Data

At the high level, the goal of the main task is to learn the embedding of different units such that the behavioral data in the corpus $\mathcal{C}$ can be reconstructed as much as possible. Our reconstruction model for the main task is inspired by the Skip-Gram model (Mikolov et al. 2013). Specifically, given a record $r \in \mathcal{C}$, let i and j be any two units that appear in r (i and j could be spatial, temporal, or textual). Then we model the probability of observing j based on unit i as:

$$p(j|i) = 1/(1 + \exp(-\mathbf{v}_i^{\mathrm{T}} \cdot \mathbf{v}_j), \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the embeddings for unit i and j, respectively. Equation 1 defines the probability of reconstructing unit j based on unit i. The overall loss function for $\mathcal{C}$ is then defined as follows:

$$O_{\mathcal{C}} = -\sum_{r \in \mathcal{C}} \sum_{i,j \in r} \log p(j|i). \tag{2}$$

Auxiliary Tasks: Preserving Correlation Graphs

Optimizing the loss function in Equation 2 is essentially leveraging the co-occurrence statistics in the corpus $\mathcal{C}$ to obtain the embeddings of different units. However, when the corpus $\mathcal{C}$ is small, the co-occurrence information may be inadequate and the learned embeddings may generalize poorly. To address this issue, we design auxiliary tasks based on the external sources $S_1, \ldots, S_N$.

Given an external source S_n, let G_n represent the correlation graph of the units extracted from S_n. The auxiliary task for S_n is designed to preserve the structure of graph G_n, such that two units that are close to each other in G_n are encouraged to have close embeddings. In specific, for any unit i, let $\mathbf{v}_i$ be the main embedding of i. In addition to $\mathbf{v}_i$, we also introduce an auxiliary embedding of i for the considered auxiliary task, denoted as $\mathbf{v}'_i$. Then we concatenate $\mathbf{v}_i$ and $\mathbf{v}'_i$ to derive the extended embedding of unit i, namely:

$$\hat{\mathbf{v}}_i = \mathbf{v}_i \oplus \mathbf{v}'_i.$$

The rationale behind the introducing the auxiliary embedding $\mathbf{v}'_i$ is that, $\mathbf{v}'_i$ has the potential to *annul the discrepancy between the main task and the auxiliary task*. Even if the correlation structures among the units are different for the main task and the auxiliary one, the auxiliary embedding $\mathbf{v}'_i$ can capture the task-specific variations of the embeddings, thus allowing the main embeddings to have more freedom to align with the common correlation structures shared by the two tasks.

After obtaining the extended embedding for every unit i, our goal is to optimize the representations such that the structure of graph G_n is well encoded. Specifically, we model the neighborhood distribution of each node based on the extended embeddings, and encourage such embedding-based distributions align well with the true distributions observed in G_n. Consider a node i with node type X and a node j with node type Y. Based on the extended embeddings, we model the likelihood of observing j given i as

$$\hat{p}(j|i) = \exp(\hat{\mathbf{v}}_j^{\mathrm{T}} \cdot \hat{\mathbf{v}}_i) \Big/ \sum_{k \in Y} \exp(\hat{\mathbf{v}}_k^{\mathrm{T}} \cdot \hat{\mathbf{v}}_i). \tag{3}$$

Equation 3 specifies the embedding-based distribution for node i, and the true observed distribution of i is defined as:

$$p(j|i) = w_{ij}/d_i \quad (4)$$

where w_{ij} is the weight of the edge e_{ij}, and $d_i = \sum_{j' \in Y} w_{ij'}$ is the total out-degree of node i for type Y.

Based on Equation 3 and 4, we define the following objective function to preserve the subgraph structure in G_n for node types X and Y:

$$O_{XY} = \sum_{i \in X} d_i \mathrm{D}(p'(\cdot|i)||p(\cdot|i)) + \sum_{j \in Y} d_j \mathrm{D}(p'(\cdot|j)||p(\cdot|j)),$$

where $\mathrm{D}(\cdot)$ is the KL-divergence measure. By minimizing O_{XY}, we are encouraging the embedding-based distributions close to the observed distributions for data types X and Y. Note that the graph can contain three different data types (location, time, text), we define the overall loss functions for preserving the graph structure as:

$$O_{G_n} = O_{WW} + O_{LL} + O_{TT} + O_{WL} + O_{WT} + O_{LT}. \quad (5)$$

Optimization

In the above, Equation 2 specifies the loss function for reconstructing the behavioral data in $\mathcal{C}$, and Equation 5 gives the loss function for preserving the correlation graph extracted from each source S_n. Now we combine the main task and the N auxiliary tasks to derive the overall loss function:

$$O = O_{\mathcal{C}} + \sum_{n=1}^{N} \lambda_n O_{G_n},$$

where $\lambda_n > 0 (1 \leq n \leq N)$ are pre-defined parameters for controlling the weights of different auxiliary sources.

To optimize the above objective function, we use negative sampling (Mikolov et al. 2013) and Adam (Kingma and Ba 2014) for efficient updating. At each time, we randomly select one of the $N + 1$ tasks (one main task and N auxiliary tasks) according to the weights of different tasks. If the main task is selected, we randomly sample a record r from $\mathcal{C}$ and a unit $i \in r$. Further, we select K random negative units that have the same type with i but do not appear in r. Then we minimize the following function for the selected samples:

$$O_r = -\log \sigma(s(i, r_{-i})) - \sum_{k=1}^{K} \log \sigma(-s(k, r_{-i})),$$

where $\sigma(\cdot)$ is the sigmoid function. We can obtain the updating rules for different variables by taking the derivatives of the above objective and then applying gradient descent using Adam (Kingma and Ba 2014).

Meanwhile, if an auxiliary task S_n is selected, we first randomly sample an edge e_{ij} and then K nodes that do not connect to node i. We consider node j as a positive example, and the K nodes as negative examples, then minimize the following function:

$$O_e = -\log \sigma(\hat{\mathbf{v}}_j^{\mathrm{T}} \cdot \hat{\mathbf{v}}_i) - \sum_{k=1}^{K} \log \sigma(-\hat{\mathbf{v}}_k^{\mathrm{T}} \cdot \hat{\mathbf{v}}_i).$$

Again, the updating rules for different embeddings (both the main embedding and the auxiliary embedding) can be easily derived by taking the derivatives of the above objective and applying Adam.

Experiments

In this section, we study the empirical performance of our proposed BRANCHNET model. We first describe our experimental setup, then report and discuss about the experimental results. We implemented our model and other baselines with Tensorflow[1], and conducted the experiments on a machine with Intel Xeon 2.80GHz CPU using 20 threads.

Experimental Setup

Data In our empirical evaluation, each experimental run requires two sets of data: (1) the target user behavioral data $\mathcal{C}$; and (2) the transferring sources $S_1, \ldots, S_N$.

1. We use the geo-tagged social media data from (Zhang et al. 2017b) as target user behavioral data. Since we focus on how effectively our proposed method can transfer external sources for *small-size* user behavioral data, we extract the following two subsets from the original datasets:
   - LA: The first extracted behavioral data contains the geo-tagged tweets created in Los Angeles. From the original Tweet dataset in (Zhang et al. 2017b), we extracted the geo-tagged tweets created during 2014.09.01 and 2014.09.20, which consists of 194,353 geo-tagged tweets. We partition the Los Angeles area into 100 * 100 equal-size grids and consider each grid as a basic spatial unit.
   - NY: The second user behavioral dataset is extracted from the original 4SQ dataset in (Zhang et al. 2017b). We extracted the Foursquare checkins created in the New York City during 2011.02.01 - 2011.05.01, which results in 31,343 Foursquare checkins. Similarly, we partition the New York City area into 100 * 100 grids to handle spatial continuity.
2. Our used transferring sources include the following:
   - **WordNet:** WordNet is a lexical database of English, which groups English words (nouns, verbs, adjectives, *etc.*) into synonyms. Given a user behavior corpus $\mathcal{C}$, we extract the keywords in $\mathcal{C}$ that appear in the WordNet database, and construct an unweighted graph for those keywords. There exists an edge between two keywords if they are synonyms in WordNet.
   - **OtherCity:** For each of the two behavioral datasets, LA and NY, we also attempt to transfer knowledge about the keywords from other cities. To augment LA, we construct a correlation graph for the keywords using the check-in data in New York City. Specifically, we first extract the keywords in LA as the graph nodes, then we take the Foursquare check-ins in NYC during 2014.08.01 – 2014.08.30. For any two keywords in LA, we connect them with an undirected edge if they co-occurred in the same check-in, and set the edge weight to the number of co-occurrences. Similarly, to augment NY, we use the geo-tagged tweets in LA during 2014-08.01 – 2014.08.30 and build a keyword co-occurrence graph for the keywords in NY.

[1] https://www.tensorflow.org/

- **SameCity:** The third transferring source are the user behavioral data in the same city but during different time periods. To augment LA, we take the geo-tagged tweets created during 2014.08.01 – 2014.08.30, and use those tweets to build a co-occurrence correlation graph for the spatial, temporal, and textual units in LA. For any two units, we connect them with an edge if they have co-occurred, then we set the edge weight to the normalized number of co-occurrences. For NY, we take the check-ins in NYC during 2010.08.01 – 2011.01.30 and construct a co-occurrence graph.

Baselines We compare BRANCHNET with the following methods:

- CROSSMAP (Zhang et al. 2017b) is a state-of-the-art method for spatiotemporal activity modeling. Given a behavioral data set $\mathcal{C}$, it first detects spatial and temporal hotspots using the mean shift algorithm, and then derives cross-modal embedding for the spatial, temporal, and textual units. However, CROSSMAP is unable to transfer knowledge from external sources.
- FINETUNE is another baseline based on fine tuning. Given the external sources $S_1, \ldots, S_N$, it first sequentially consumes these sources to learn embeddings for different units based on graph embedding (Tang et al. 2015). Then given the corpus $\mathcal{C}$, it treats the learned embeddings as initializations, and fine tune the embeddings to reconstruct user behaviors in $\mathcal{C}$.
- SEMIEMBED (Weston, Ratle, and Collobert 2008) is a semi-supervised method based on graph regularization. Given a set of observed units, it passes the embeddings into a deep neural network for predicting the target unit. To leverage the correlation graph, it incorporates a regularization term such that the embeddings of two terms that are connected in the graph tend to have close embeddings.
- PLANETOID (Yang, Cohen, and Salakhutdinov 2016) is also a multi-task framework for graph-based semi-supervised learning. It assumes the embeddings of the units are shared for both the main prediction task as well as the task of preserving graph structures. The major difference between PLANETOID and BRANCHNET is that, we design task-specific embeddings for different auxiliary tasks to capture the discrepancies among different tasks.

Parameter Settings In our experiments, we use the main embedding dimension to 400 for all the methods by default, and set the task-specific embedding dimensions to 100. When using Adam to learn the embeddings, we set the learning rate to 0.002 and train for 10 epochs. The methods of SEMIEMBED, PLANETOID, and BRANCHNET require transferring knowledge from external sources, and we set the default weight for an auxiliary task λ_n to 0.1.

Evaluation Protocol We use the activity reconstruction task to evaluate performance of different models. Given a corpus $\mathcal{C}$ (*i.e.*, LA or NY), we randomly split $\mathcal{C}$ into two different subsets: 80% for model training, and 20% for test. Recall that each record r has three attributes: location, time, and text. We thus have three reconstruction tasks in total: (1) predicting the location based on the given time and text; (2) predicting the time based on the given location and text; and (3) predicting the text message based on the given location and time. Take the location reconstruction task as an example. For the ground-truth location l_r, we mix it with a set of randomly sampled candidate locations. Then we use the observed time t_r and text message w_r to rank the candidate locations by similarity and identify the most similar one. The similarity score is computed by averaging the cosine similarities between the embedding of the candidate location and the embeddings of the observed units.

Intuitively, the better an activity model captures the cross-modal correlations between location, time, and text, the more likely it ranks the ground truth location to top positions. Following the evaluation protocol in (Zhang et al. 2017b), we use the mean reciprocal rank (MRR) to quantify the performance of different methods. Given a set Q of queries, the MRR is computed as:

$$MRR = \frac{1}{|Q|} \sum_{i}^{Q} \frac{1}{\text{rank}_i}, \tag{6}$$

where rank_i is the ranking of the ground truth for the i-th query.

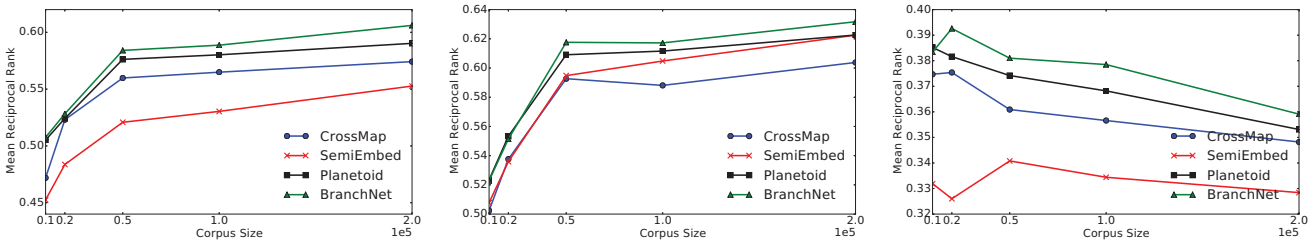
Experimental Results

Performance Comparison with Baseline Methods Table 1 shows the results of all the methods for activity reconstructions on LA and NY. As shown, BRANCHNET outperforms the baseline methods in all the three types of prediction tasks. BRANCHNET improves the performance of CROSSMAP by as much as 5.9%, which shows that BRANCHNET is able to effectively transfer knowledge from the three external sources to alleviate the data scarcity problem. PLANETOID is the strongest among the baseline methods, but BRANCHNET consistently outperforms PLANETOID in different settings. The reason is BRANCHNET includes branched task-specific embeddings, which better cope with the discrepancies between the main task and the auxiliary tasks.

Method	Location		Text		Time	
	Tweet	4SQ	Tweet	4SQ	Tweet	4SQ
CROSSMAP	0.5733	0.5270	0.5892	0.5427	0.3609	0.3607
FINETUNE	0.5787	0.5261	0.5929	0.5455	0.3623	0.3615
SEMIEMBED	0.5100	0.4876	0.5968	0.5520	0.3354	0.3371
PLANETOID	0.5812	0.5505	0.6163	0.5634	0.3639	0.3721
BRANCHNET	**0.5904**	**0.5567**	**0.6241**	**0.5659**	**0.3730**	**0.3764**

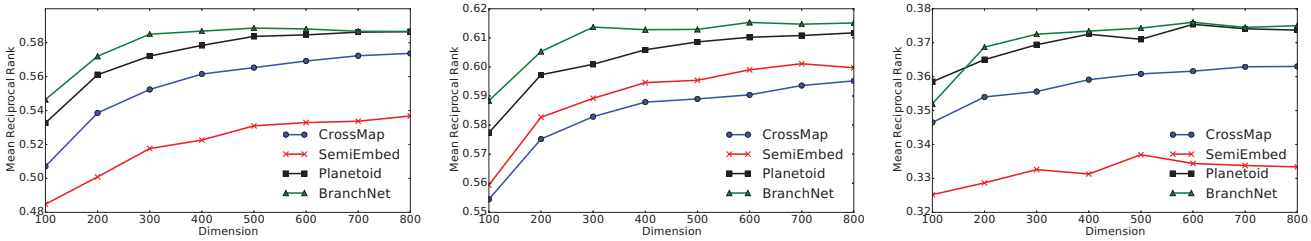
Table 1: The MRRs of different methods for activity reconstruction.

Performance on Different-Size Corpora In this set of experiments, we study the effectiveness for transferring external knowledge for user behavioral corpora with different sizes. For this purpose, we take the LA dataset and downsample it to generate multiple subsets with different sizes (10 thousand, 50 thousand, and 100 thousand). Using the



(a) MRR for location prediction. (b) MRR for text prediction. (c) MRR for time prediction.

Figure 3: The transferring effectiveness of different methods on the corpora with different sizes.



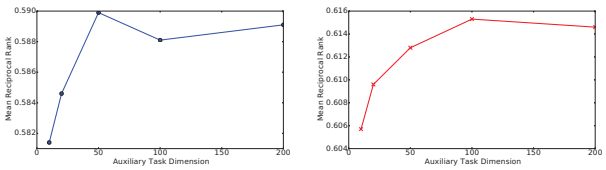
(a) MRR for location prediction. (b) MRR for text prediction. (c) MRR for time prediction.

Figure 4: The performance of different methods when the dimension of the main embedding varies.

same transferring sources described earlier, we apply different methods for those different-size user behavioral corpora and evaluate the performance for spatiotemporal activity prediction. Figure 3 shows the performance when the size of user behavior corpus increases. As shown, with an increasing corpus size, the absolute performance of all the methods for spatiotemporal activity prediction increases accordingly. This is reasonable since the larger the dataset, the more information it contains, which helps learn more reliable spatiotemporal activity models.

Among all the methods, BRANCHNET still achieves the best performance for the corpora with different sizes. Such a phenomenon shows the branched multi-task embedding structure is quite robust. Comparing the performance between CROSSMAP and BRANCHNET for different corpus sizes, we find the relative performance gain is particularly large when the corpus size is small. For example, when the data size is 10 thousand, BRANCHNET improves the performance of location prediction by 3%. This is because when the user behavioral corpus $\mathcal{C}$ is scarce, the usefulness of the knowledge from external sources is more evident. Another interesting finding is that, while BRANCHNET is generally superior to PLANETOID, the performance gap between them is not large on small-size corpus. The reason is probably that, when the user behavioral corpus is too small (*e.g.*, 10 thousand), there is not enough evidence for BRANCHNET to discriminate between the embedding structures that should be shared by different tasks and the embedding structures that are task-specific. However, as the size of the user behavioral corpus increases, the performance gap between BRANCHNET and PLANETOID becomes obvious.

Effects of the Embedding Dimensionality We finally study the effects of the embedding dimensionality on the performance of different methods. Figure 4 shows the results when we vary the dimensionality of the main embedding from 100 to 800. As shown, the performance of all the methods first increases with the dimension size, and then gradually stabilizes after the dimensionality is larger than 400. This phenomenon is intuitive, because a larger embedding dimensionality leads to spatiotemporal activity models that have better expressive power. On the other hand, a too large dimensionality makes the model harder to learn and incurs more computational cost. That is why we set the main embedding dimension to 400 by default. Figure 5 shows the effect of the task-specific embedding dimensionality on the performance of BRANCHNET for location prediction and text prediction (we omit the plot for the time prediction due to the space limit). As shown, when we vary the dimensionality of the auxiliary embedding from 10 to 200, the performance first increases and then stabilizes after 100.



(a) MRR for location prediction. (b) MRR for text prediction.

Figure 5: The effects of the auxiliary embedding dimensionality.

Related Work

In this section, we review existing work related to our problem. We describe relevant works from the following areas: spatiotemporal activity modeling, graph-based semi-supervised learning, and transfer learning.

Spatiotemporal activity modeling. Existing spatiotemporal activity modeling methods can be categorized into two classes: *topic-model-based* (Sizov 2010; Kling et al. 2014; Yin et al. 2011; Hong et al. 2012; Yuan et al. 2013) and *embedding-based* (Xie et al. 2016; Zhang et al. 2017b; 2017c). Generally, the former extends classic topic models to bridge different data modalities, by assuming each latent topic can generate observations over not only textual keywords but also locations and timestamps. For example, Sizov *et al.* (Sizov 2010) extend LDA (Blei, Ng, and Jordan 2003) by assuming each latent topic has a multinomial distribution over text, and two Gaussians over latitudes and longitudes. They later extend the model to find topics that have complex and non-Gaussian distributions (Kling et al. 2014). Yin *et al.* (Yin et al. 2011) extend PLSA (Hofmann 1999) by assuming each region has a normal distribution that generates locations, as well as a multinomial distribution over the latent topics that generates text.

One drawback of such topic-model-based methods is that they have to impose distribution assumptions on different modalities, which may not fit the true distributions in the real data well. To address this problem, embedding-based methods have been recently proposed. (Zheng et al. 2012) *et al.* build a user-location-activity tensor and use factorization to learn latent representations for users and locations for personalized recommendation. Zhang *et al.* (Zhang et al. 2017b) first detect spatial and temporal hotspots where people's activities burst, and then map different regions, hours, and activities into the same latent space such that correlated units tend to have close embeddings. They later propose a semi-supervised cross-modal embedding by incorporating activity category information (Zhang et al. 2017c). Our work is quite related to (Zhang et al. 2017b; 2017c) as we also use cross-modal embedding for spatiotemporal activity modeling. However, they do not consider the data scarcity problem and may produce poor performance with small user behavior data. In contrast, we aim to design effective approaches that effectively transfer the knowledge from external sources to alleviate data scarcity.

Graph-Based Semi-supervised Learning. For each external source, we use a heterogeneous graph to encode the correlations between different units for knowledge transferring. Such an idea is closely related to existing works on graph-based semi-supervised learning. Given a target prediction task that is short of training data, graph-based semi-supervised learning aims to leverage the information encoded in a graph to augment the target task. Conventional graph-based semi-supervised learning designs an objective function that consists of both: 1) the loss function over the target task; and 2) the regularization term for preserving the graph structure. Different methods have been proposed under this framework and they mainly differ in terms of how they preserve the graph structure, and representative methods include Gaussian Random Fields (Zhu, Ghahramani, and Lafferty 2003), smoothness constraint (Zhou et al. 2003), and manifold regularization (Belkin, Niyogi, and Sindhwani 2006).

Besides these traditional graph regularization approaches, recent years have witnessed growing interest in using graphs to guide representation learning (Weston, Ratle, and Collobert 2008; Yang, Cohen, and Salakhutdinov 2016; Yang et al. 2017). For instance, Weston *et al.* (Weston, Ratle, and Collobert 2008) impose graph regularization into the embedding learning process, such that two instances that are close to each other in the graph are encouraged to have close embeddings. Yang *et al.* (Yang et al. 2017) combine graph regularization with collaborative filtering for effective POI recommendation. The most similar model to ours is (Yang, Cohen, and Salakhutdinov 2016), which also adopts a multi-task learning framework, and the embeddings are jointly learned for a prediction task over the target domain as well as a graph embedding task that preserves graph structures. The major difference between our model and (Yang, Cohen, and Salakhutdinov 2016) is that we design task-specific branches in our model, such that the discrepancies across different tasks are captured.

Transfer Learning There is a large body of literature on transfer learning (Cao et al. 2010; Long et al. 2012; Pan and Yang 2010; Long et al. 2017), which aims at transferring the knowledge from a source domain to a target domain. The difference between transfer learning and our work is two-fold. First, in transfer learning, the classifier for the source domain is typically learned beforehand, and the focus is to adapt the knowledge in the classifier for the target domain (Pan and Yang 2010). In contrast, in our BRANCHNET model, the embeddings for the source and target domains are learned simultaneously. Second, transfer learning typically assumes the data from the source domain is labeled and shares the same label space with the target domain, whereas we do not have such assumptions, but use a flexible graph to encode the general knowledge from external sources and learn the embeddings in an unsupervised way.

Conclusion

We have studied the problem of modeling people's spatiotemporal activities from a limited amount of user behavior data. We proposed a graph-regularized cross-modal embedding framework. It uses heterogeneous graphs to encode information from external sources, and then employs a multi-task framework to learn the cross-modal embeddings of location, time, and text. It features a main embedding space that learns the correlation structures shared by the main task and auxiliary tasks, as well as auxiliary embedding spaces that captures the discrepancies among tasks. Our experiments show that the proposed model can effectively transfer information from external sources to help spatiotemporal activity modeling, and it outperforms state-of-the-art graph-regularized semi-supervised methods.

Acknowledgements

This work was sponsored in part by the U.S. Army Research Lab under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and Grant 1U54GM114838 awarded by NIGMS. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing any funding agencies.

References

Abdelhaq, H.; Sengstock, C.; and Gertz, M. 2013. Eventweet: Online localized event detection from twitter. *PVLDB* 6(12):1326–1329.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(1):993–1022.

Cao, B.; Pan, S. J.; Zhang, Y.; Yeung, D.; and Yang, Q. 2010. Adaptive transfer learning. In *AAAI*.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring millions of footprints in location sharing services. In *ICWSM*, 81–88.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, 58 – 65.

Feng, W.; Zhang, C.; Zhang, W.; Han, J.; Wang, J.; Aggarwal, C. C.; and Huang, J. 2015. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *ICDE*, 1561–1572.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A. J.; and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the twitter stream. In *WWW*, 769–778.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Kling, C. C.; Kunegis, J.; Sizov, S.; and Staab, S. 2014. Detecting non-gaussian geographical topics in tagged photo collections. In *WSDM*, 603–612.

Long, M.; Wang, J.; Ding, G.; Shen, D.; and Yang, Q. 2012. Transfer learning with graph co-regularization. In *AAAI*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*, 2208–2217.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *CoRR* abs/1310.4546.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–1359.

Sizov, S. 2010. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM*, 281–290.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. LINE: large-scale information network embedding. *CoRR* abs/1503.03578.

Weston, J.; Ratle, F.; and Collobert, R. 2008. Deep learning via semi-supervised embedding. In *ICML*, 1168–1175.

Xie, M.; Yin, H.; Wang, H.; Xu, F.; Chen, W.; and Wang, S. 2016. Learning graph-based POI embedding for location-based recommendation. In *CIKM*, 15–24.

Yang, C.; Bai, L.; Zhang, C.; Yuan, Q.; and Han, J. 2017. Bridging collaborative filtering and semi-supervised learning: A neural approach for POI recommendation. In *KDD*, 1245–1254.

Yang, Z.; Cohen, W. W.; and Salakhutdinov, R. 2016. Revisiting semi-supervised learning with graph embeddings. In *ICML*, 40–48.

Yao, S.; Hu, S.; Zhao, Y.; Zhang, A.; and Abdelzaher, T. F. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *WWW*, 351–360.

Yin, Z.; Cao, L.; Han, J.; Zhai, C.; and Huang, T. S. 2011. Geographical topic discovery and comparison. In *WWW*, 247–256.

Yuan, Q.; Cong, G.; Ma, Z.; Sun, A.; and Thalmann, N. M. 2013. Who, where, when and what: discover spatio-temporal topics for twitter users. In *KDD*, 605–613.

Zhang, C.; Zhou, G.; Yuan, Q.; Zhuang, H.; Zheng, Y.; Kaplan, L. M.; Wang, S.; and Han, J. 2016. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *SIGIR*, 513–522.

Zhang, C.; Liu, L.; Lei, D.; Yuan, Q.; Zhuang, H.; Hanratty, T.; and Han, J. 2017a. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *KDD*, 595–604.

Zhang, C.; Zhang, K.; Yuan, Q.; Peng, H.; Zheng, Y.; Hanratty, T.; Wang, S.; and Han, J. 2017b. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *WWW*, 361–370.

Zhang, C.; Zhang, K.; Yuan, Q.; Tao, F.; Zhang, L.; Hanratty, T.; and Han, J. 2017c. React: Online multimodal embedding for recency-aware spatiotemporal activity modeling. In *SIGIR*, 245–254.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2012. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. *Artif. Intell.* 184-185:17–37.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. In *NIPS*, 321–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. D. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 912–919.