Community-Aware Multi-Task Transportation Demand Prediction

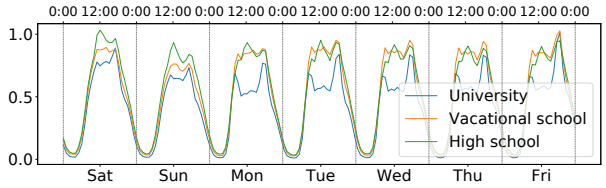
Hao Liu[1*†], Qiyu Wu[1,2†], Fuzhen Zhuang[3,4], Xinjiang Lu[1], Dejing Dou[1], Hui Xiong[5]
[1]Baidu Research, Beijing, China
[2]Peking University, China
[3]Key Lab of IIP of Chinese Academy of Sciences (CAS), ICT, CAS, Beijing 100190, China
[4]University of Chinese Academy of Sciences, Beijing 100049, China
[5]Rutgers University, USA
{liuhao30, luxinjiang, doudejing}@baidu.com, wuqiyu@pku.edu.cn, zhuangfuzhen@ict.ac.cn, hxiong@rutgers.edu

Abstract

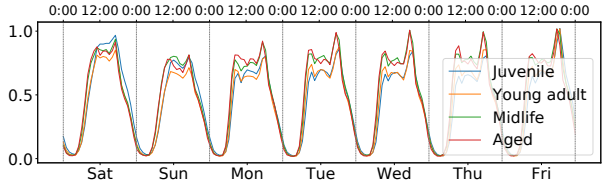
Transportation demand prediction is of great importance to urban governance and has become an essential function in many online applications. While many efforts have been made for regional transportation demand prediction, predicting the diversified transportation demand for different communities (e.g., the aged, the juveniles) remains an unexplored problem. However, this task is challenging because of the joint influence of spatio-temporal correlation among regions and implicit correlation among different communities. To this end, in this paper, we propose the ***M**ulti-task **S**patio-**T**emporal **Net**work with Mutually-supervised **Ada**ptive task grouping* (**Ada-MSTNet**) for community-aware transportation demand prediction. Specifically, we first construct a sequence of multi-view graphs from both spatial and community perspectives, and devise a spatio-temporal neural network to simultaneously capture the sophisticated correlations between regions and communities, respectively. Then, we propose an adaptively clustered multi-task learning module, where the prediction of each region-community specific transportation demand is regarded as distinct task. Moreover, a mutually supervised adaptive task grouping strategy is introduced to softly cluster each task into different task groups, by leveraging the supervision signal from one another graph view. In such a way, Ada-MSTNet is not only able to share common knowledge among highly related communities and regions, but also shield the noise from unrelated tasks in an end-to-end fashion. Finally, extensive experiments on two real-world datasets demonstrate the effectiveness of our approach compared with seven baselines.

Introduction

Urban transportation demand prediction aims at forecasting the amount of crowd who intend to get in or leave out of regions in a city in the next time period. Due to its importance to urban governance and commercial applications (Moreira-Matias et al. 2013), many efforts have been made for transportation demand prediction. However, after analyzing real-world mobility data, we find the transportation demand of different communities is diversified yet correlated. For example, as shown in Figure 1(a), the communities in different education levels show different demand patterns with a strong temporal periodicity. In particular, we observe a highly synchronized demand pattern between high school and vocational school communities, but a diverged demand pattern of University students. Similar observations can also be found between communities in different age level, as shown in Figure 1(b). Undoubtedly, predicting community level transportation demand is valuable for fine-grained public service and business operation. For example, it can help dynamically rebalancing shared-bikes for students, dispatching taxis at midnight for females, and pre-allocate more high-end vehicles to regions where high-consumption communities have demands. In this work, we study the problem of *Community-Aware Transportation Demand Prediction* (**CATDP**), where the regional transportation demands of different communities (e.g., the aged, the juveniles) are predicted simultaneously.



(a) Transportation demand versus different education levels.



(b) Transportation demand versus different age levels.

Figure 1: Real-world transportation demand patterns of representative communities in a downtown area in Beijing, ranged from June 22, 2019 to June 28, 2019.

However, three major challenges arise toward CATDP. First, the transportation demand in different regions is both spatially and temporally correlated. The transportation demand of a region is not only influenced by its adjacent regions but also correlated with its transportation demand in past time periods. How to model the spatio-temporal au-

*Corresponding author.
†Equal contribution.

tocorrelation is the first challenge. Second, the transportation demand of different communities is diversified but correlated, as reported in Figure 1. Independently training existing transportation demand prediction models (Yao et al. 2018) for different communities overlooks the inner-connection among various communities. As a result, how to incorporate the diversified but synchronized transportation demand relationship between communities for CATDP is another challenge. Last, the relationships between different communities are also varying under different spatio-temporal context. Simply regarding CATDP as a multivariate time-series prediction task and seamlessly sharing information among all communities in different regions may introduce unexpected noises and degenerate the prediction performance (Yao, Cao, and Chen 2019). How to selectively share knowledge between highly correlated region-community specific sub-tasks is the third challenge.

In order to tackle the above challenges, we propose the ***M**ulti-task **S**patio-**T**emporal **Net**work with Mutually-supervised **Ada**ptive task grouping* (**Ada-MSTNet**) for CATDP. First, to characterize community-aware spatio-temporal correlations, we construct a sequence of multi-view graphs from both the spatial perspective and the community perspective. Then we introduce a spatio-temporal neural network, consisting of a Graph Neural Network (GNN) block and a Recurrent Neural Network (RNN) block for spatial dependency and temporal dependency modeling, respectively. After that, we introduce a soft clustered multi-task learning module, where the transportation demand prediction of each community in different regions is regarded as distinct task. Moreover, to selectively share information among related tasks and prevent noise diffusion across unrelated tasks, we further propose an adaptive task grouping strategy to dynamically cluster tasks into different task groups, which is supervised by the region or community representation obtained from one another graph view. In such a way, Ada-MSTNet is not only able to capture sophisticated spatio-temporal and community correlations, but also selectively share common knowledge among highly related communities and regions for transportation demand prediction. Finally, extensive experiments on real-world datasets demonstrate the effectiveness of the proposed model compared with state-of-the-art solutions.

Preliminaries

We first split the entire city into an $a \times b = n$ non-overlapping grid map, which consist of a rows and b columns, denoted by $\mathcal{R} = \{r_1, r_2, \dots, r_n\}$. Besides, we split the whole time period (*e.g.*, two months) into T equal-length time intervals, denoted by $\mathcal{T} = \{t_1, t_2, \dots, t_T\}$.

Definition 1. ***Community***. *A community is defined as a crowd of individuals sharing a common identity, such as age, gender, and education level, etc.. Considering a set of identity attributes $A = \{a_1, a_2, \dots\}$, we can define m corresponding communities, denoted by $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$.*

For each individual, we can derive identities based on static user profiles or dynamic user interests. For example, an individual can be attributed as a man or woman based on gender. On the one hand, each individual may have numerous attributes and therefore can belong to multiple communities. On the other hand, each community is often partially overlapping, and the total sum of the crowd in each community is greater than the overall population.

Definition 2. ***Query***. *A query is defined as a four-tuple $q = (u, o, d, t)$, where u is an individual, o is the origin region, d is the destination region, and t is the departure time interval.*

In online applications such as navigation tool or ride-hailing platform, a query indicates a transportation demand of u move from o to d at time t.

Definition 3. ***Transportation demand***. *Based on the movement direction, the transportation demand can be either get in or leave out. For a region r_i and a community c_j, the inflow and outflow transportation demand at time interval t are respectively defined as the query volume of c_j origin from r_i and destination at r_i, denoted by $x^t_{out,i,j}$ and $x^t_{in,i,j}$.*

At a specific time interval t, the transportation demand can be denoted by a tensor $\mathbf{X}^t \in \mathbb{R}^{2\times N\times M}$, in which $\mathbf{X}^t_{in} \in \mathbb{R}^{N\times M}$ and $\mathbf{X}^t_{out} \in \mathbb{R}^{N\times M}$ correspond to inflow and outflow demand, $\mathbf{X}^t_i \in \mathbb{R}^{2\times M}$ is the demand of all communities get in or leave out region r_i, and $\mathbf{X}^t_j \in \mathbb{R}^{2\times N}$ is the demand of community c_j get in or leave out all regions. The demand of r_i and c_j is denoted by $\mathbf{X}^t_{i,j}$, and we stipulate the first subscript indicates region index when both indices occur. Note that at a specific time interval, the overall inflow demand equals the overall outflow demand, *i.e.*, $\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{X}^t_{in,i,j} = \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{X}^t_{out,i,j}$.

Problem 1. ***Community-aware transportation demand prediction (CATDP)***. *Given T historical time intervals, and observed historical transportation demands $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^T\}$, our problem is to predict transportation demands for next τ time intervals,*

$$\mathcal{F}(\boldsymbol{\mathcal{X}}) \rightarrow (\mathbf{X}^{T+1}, \mathbf{X}^{T+2}, \dots, \mathbf{X}^{T+\tau}), \tag{1}$$

where $\mathcal{F}(\cdot)$ is the mapping function we aim to learn.

Framework Overview

Figure 2 shows an overview of Ada-MSTNet, which including the following three major tasks: (1) the construction of time-dependent multi-view transportation demand graphs; (2) the spatio-temporal and community correlation modeling; and (3) the selective knowledge sharing between highly correlated communities in different spatio-temporally correlated regions. In the first task, we split the historical transportation demands into a sequence of graph snapshots over time, and construct time-dependent multi-view graphs based on spatial adjacency (*i.e.*, the region view) and historical flow sequence similarity (*i.e.*, the community view). In the second task, we devise a *spatio-temporal neural network* on both views to simultaneously capture the spatio-temporal and community correlation. In the third task, the city-wide community-aware transportation demand is obtained via an *adaptively clustered multi-task learning* component with mutually supervised adaptive task grouping for end-to-end task group generation.

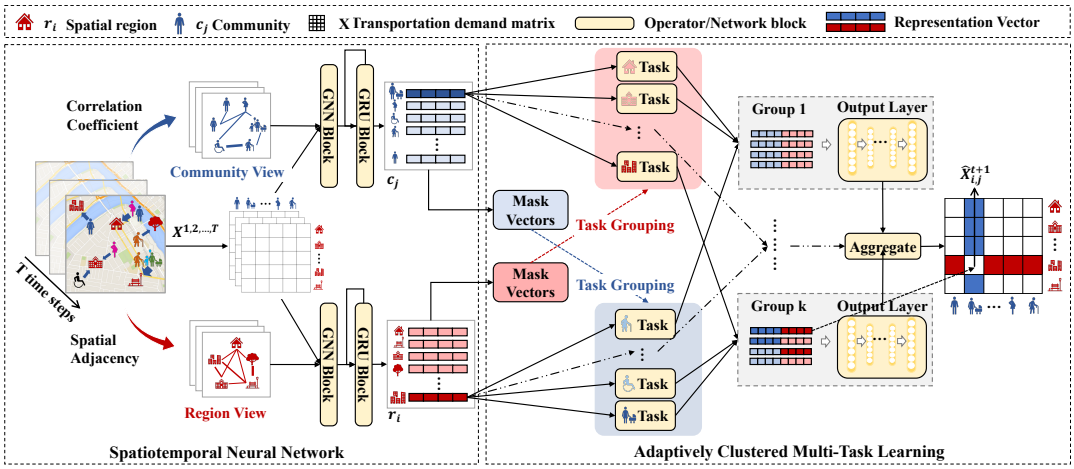


Figure 2: The framework overview of Ada-MSTNet.

Transportation Demand Graphs Construction

We construct time-dependent multi-view graphs of city transportation demand from two perspectives: (1) region spatial adjacency, and (2) community demand similarity.

The region graph view. The region graph demonstrates the geographical connectivity among regions according to spatial adjacency. Specifically, we construct the region graph $\mathcal{G}^r = (\mathcal{R}, \mathcal{E}^r, \mathcal{A}^r)$, where $\mathcal{R}$ is the set of regions, $\mathcal{E}^r$ is the the set of edges between regions, and $\mathcal{A}^r$ denotes the proximity matrix of $\mathcal{G}^r$ which will be learned by our model automatically. For two regions r_i and r_j, we define their connectivity as

$$e_{i,j} = \begin{cases} 1, & adj(r_i, r_j) \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $adj(\cdot,\cdot)$ is the adjacency function equals to one if and only if r_i and r_j are spatially bordered on each other.

The community graph view. The community graph captures the similarity of transportation demand between communities. We construct the community graph $\mathcal{G}^c = (\mathcal{C}, \mathcal{E}^c, \mathcal{A}^c)$, where $\mathcal{C}$ is the set of communities, $\mathcal{E}^c$ is the set of edges between communities, and $\mathcal{A}^c$ is the proximity matrix of $\mathcal{G}^c$ which is decided by our model during training. For two communities c_i and c_j, we define their connectivity as

$$e_{i,j} = \begin{cases} 1, & pcc(c_i, c_j) \geq \epsilon \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $pcc(\cdot,\cdot)$ is the Pearson Correlation Coefficient (PCC) (Benesty et al. 2009), which measures the strength of the linear correlation between c_i and c_j. ϵ is a correlation threshold. The PCC score between c_i and c_j is defined as

$$pcc(c_i, c_j) = \frac{1}{|\mathcal{R}|} \sum_{r=1}^{|\mathcal{R}|} \frac{\sum_{t=1}^{T} (\mathbf{X}_{r,i}^t - \overline{\mathbf{X}}_{r,i})(\mathbf{X}_{r,j}^t - \overline{\mathbf{X}}_{r,j})}{\sqrt{\sum_{t=1}^{T} (\mathbf{X}_{r,i}^t - \overline{\mathbf{X}}_{r,i})^2} \sqrt{\sum_{t=1}^{T} (\mathbf{X}_{r,j}^t - \overline{\mathbf{X}}_{r,j})^2}}, \tag{4}$$

where $\overline{\mathbf{X}}_{r,i}$ and $\overline{\mathbf{X}}_{r,j}$ are averaged values of $\mathbf{X}_{r,i}^t$ and $\mathbf{X}_{r,j}^t$ over T time steps, and $\mathcal{R}$ is the set of regions.

Note that each vertex's contextual features in both region and community graphs are varying over time. For each view, we construct sequence of time-dependent graphs, $[\mathcal{G}^{r,t_1}, \mathcal{G}^{r,t_2}, \ldots, \mathcal{G}^{r,T}]$ and $[\mathcal{G}^{c,t_1}, \mathcal{G}^{c,t_2}, \ldots, \mathcal{G}^{c,T}]$, where each graph is assigning with an individual $\mathcal{A}$.

The Ada-MSTNet Model

Spatio-Temporal Neural Network

We first introduce the *spatio-temporal neural network* for simultaneous spatio-temporal dependency and community dependency modeling. Specifically, we exploit the graph neural network (GNN) (Veličković et al. 2018) to capture the structural correlation of spatial and community graphs, and devise a recurrent neural network (RNN) (Mikolov et al. 2010) for temporal correlation modeling.

Structural correlation modeling. GNN is an effective generalization of convolution neural networks for handling non-Euclidean graph structures. By iteratively aggregating one-hop neighbors and applying flexible transforming functions, GNN updates node representations with implicit local structural information preservation (Kipf and Welling 2017). Take the region graph for example, considering a region graph $\mathcal{G}^{r,t}$ at time step t, let $\mathbf{x}_i^t$ denote the current representation of region r_i, we define the graph convolution operation

$$\mathbf{x}_i^{t\prime} = \sigma\Big(\sum_{r_j \in \mathcal{N}_i} \alpha_{i,j}^t \mathbf{W}_c(\mathbf{x}_j^t || \mathbf{x}_i^t)\Big), \tag{5}$$

where σ is a non-linear activation function, $\mathbf{W}_c \in \mathbb{R}^{d \times 2d}$ is a learnable weighted matrix shared by all regions $r_i \in \mathcal{G}^{r,t}$ over all time steps, $||$ is the concatenation operation, and $\mathcal{N}_i$ is the set of regions connected with r_i in $\mathcal{G}^{r,t}$. $\alpha_{i,j}^t \in \mathcal{A}^t$ is the time and edge specific proximity score derived by

$$\alpha_{i,j}^t = \frac{\exp(\text{Attn}(\mathbf{W}_a \mathbf{x}_i^t, \mathbf{W}_a \mathbf{x}_j^t))}{\sum_{r_k \in \mathcal{N}_i} \exp(\text{Attn}(\mathbf{W}_a \mathbf{x}_i^t, \mathbf{W}_a \mathbf{x}_k^t))}, \tag{6}$$

where $\mathbf{x}_i^t$, $\mathbf{x}_j^t$ and $\mathbf{x}_k^t$ are current representations of corresponding regions, $\mathbf{W}_a \in \mathbb{R}^{d \times d}$ is learnable matrix shared by all edges in $\mathcal{G}^{r,t}$, and $Attn(\cdot,\cdot)$ is an attention function (e.g., scaled dot-product, concatenation, etc.) (Vaswani et al. 2017). Note that we learn different proximity matrix $\mathcal{A}^t$ for each time step t to model the time-varying spatial dependency among regions. The spatial correlation of community graphs is computed in the same way.

Temporal correlation modeling. Based on the learned representation of GNN, we adopt GRU (Cho et al. 2014), an effective yet more efficient variant of RNN, for temporal dependency modeling. Take the region graph again for example, consider a sequence of representation of region r_i in previous T time steps $(\mathbf{x}_i^{t-T}, \mathbf{x}_i^{t-T+1}, \ldots, \mathbf{x}_i^t)$, our model derives the status of r_i at time step t by

$$\mathbf{h}_i^t = GRU(\mathbf{h}_{i-1}^t, \mathbf{x}_i^t), \tag{7}$$

where $\mathbf{h}^t$ and $\mathbf{h}_i^{t-1}$ are respectively hidden states of r_i at time step t and t-1. Particularly, we initialize $\mathbf{h}_i^0$ by zero.

In practice, we construct the input sequence from two perspectives, *i.e.*, the closeness and periodic (Zhang, Zheng, and Qi 2017), where the closeness sequence is the T consecutive time steps before the current time step, and the periodic sequence retains the periodicity of transportation demands in a certain time interval (*i.e.*, 24 hours). Note that the temporal dependency of vertices in region graphs and in community graphs are modeled separately, as illustrated in Figure 2.

Adaptively Clustered Multi-Task Learning

Then, we present the *Adaptively clustered multi-task learning* for community-aware transportation demand prediction.

Soft clustered multi-task learning. Based on the spatio-temporal neural network, we have latent representations $\mathbf{h}_i^{r,t}$ of region $r_i \in \mathcal{R}$ and $\mathbf{h}_j^{c,t}$ of community $c_j \in \mathcal{C}$. By regarding the prediction of transportation demand of each communities in different regions as a distinct task, we formulate CADTP as a multi-task learning problem. However, as aforementioned in Figure 1, the transportation demand of each community in each region is non-uniformly correlated. Thus, we group tasks into multiple groups, where highly correlated tasks are clustered together. By enforcing tasks in the same group to share a same feature transformation function (*i.e.*, a same sub-network), common knowledge can be more effectively shared only across highly correlated tasks.

Formally, given a set of tasks $t_{i,j} \in \mathcal{T}$, where $t_{i,j}$ corresponds to the transportation demand of a community c_j at a specific region r_i, suppose we have clustered tasks into K groups, where $K < |\mathcal{T}|$, the transportation demand of task $t_{i,j}$ under clustered multi-task learning is derived by

$$(\hat{\mathbf{X}}_{i,j}^{t+1}, \hat{\mathbf{X}}_{i,j}^{t+2}, \cdots, \hat{\mathbf{X}}_{i,j}^{t+\tau}) = \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{h}_i^{r,t} || \mathbf{h}_j^{c,t}, \mathbf{M}_{i,j} \odot \theta_k), \tag{8}$$

where f_k is the corresponding prediction network of task cluster k, θ_k indicates parameters of f_k, $\mathbf{M}_{i,j}$ is the K-dimensional mask vector of task $t_{i,j}$ indicates which cluster it belongs to, and $\odot$ is element-wise product operation. When we restrict $\mathbf{M}_{i,j,k} \in \{0, 1\}$, the prediction network is strictly clustered where each task belongs to one and only one group. However, since less-correlated tasks still have partially the same coarse-grained patterns (*i.e.*, daily periodicity), it would be potentially helpful for sharing such commonality. Hereby, we set $0 \leq \mathbf{M}_{i,j,k} \leq 1$ with the constraint $\sum_{k=1}^{K} \mathbf{M}_{t_{i,j},k} = 1$, which enables probabilistic task cluster assignment in a nonuniform way.

Mutually supervised adaptive task grouping. One intermediate problem of soft clustered multi-task learning is how to decide the group assignment probability (*i.e.*, the mask tensor $\mathbf{M}$). Conventional approaches either statically decide task groups by solving a partition problem (Kang, Grauman, and Sha 2011) or simply regard each task as a linear combination of other representative tasks (Yao, Cao, and Chen 2019). Different from existing studies, we propose a *mutually supervised adaptive task grouping* strategy to adaptively cluster tasks into multiple task groups with soft assignment weights. As defined in Equation (8), the hidden state of the region r_i is shared by m tasks, and each of which corresponds to the transportation demand of a different community in r_i. The key insight of the mutually supervised strategy is that the group of r_i involved m tasks can be supervised by hidden states of each c_j obtained from the one another graph view, and vice versa.

Specifically, consider the community c_j and the corresponding hidden state $\mathbf{h}_j$, the assignment weights for K groups of c_j related tasks are derived by a Softmax classifier (Goodfellow, Bengio, and Courville 2016),

$$\mathbf{M}_{:,j}^{c} = \text{Softmax}(\mathbf{W}_g^\top \mathbf{h}_j'), \tag{9}$$

where $\mathbf{M}_{:,j}^{c}$ is the K-dimensional mask vector of c_j related tasks, $\mathbf{W}_g \in \mathbb{R}^{d \times K}$ is the learnable weighted matrix. In particular, $\mathbf{h}_j' = \sigma(\mathbf{W}_f \mathbf{h}_j)$ is the transformed community representation of c_j, where σ is a non-linear activation function and $\mathbf{W}_f \in \mathbb{R}^{d \times d}$ is the learnable parameter.

We introduce an extra self-supervised task to guarantee the transformed community representation preserves salient features of c_j for task grouping, which is implemented by optimizing $\hat{\mathbf{c}}_j = \mathbf{W}_c \mathbf{h}_j'$, where $\hat{\mathbf{c}}_j$ is the estimated community distribution h_j' belongs to. Likewise, we derive $\mathbf{M}_{:,i}^{r}$ for each region r_i, which is optimized by corresponding self-supervised tasks. Overall, the mask vector for task $t_{i,j}$ is a combination from two graph views, $\mathbf{M}_{i,j} = \frac{1}{2}(\mathbf{M}_{:,i}^{r} + \mathbf{M}_{:,j}^{c})$. In such a way, all tasks can be assigned to K groups, guided by hidden states mutually supervised from two graph views. The mask tensor $\mathbf{M}$ is learned adaptively along with the CATDP task in an end-to-end fashion rather than relaying on any predefined clustering assumption.

Training and Optimization

Ada-MSTNet aims to minimize the *mean absolute error* (MAE) between the predicted transportation demand and the observed transportation demand in next τ time steps,

$$O_m = \frac{1}{n \times m}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{\tau} |\hat{\mathbf{X}}_{i,j}^{t+k} - \mathbf{X}_{i,j}^{t+k}|. \tag{10}$$

In addition, the loss of self-supervised task in community view is defined as

$$O_c = -\frac{1}{m^2}\sum_{i}^{m}\sum_{j}^{m} \mathbf{c}_j^{(i)} \log \frac{exp(\hat{\mathbf{c}}_j^{(i)})}{\sum_k^m exp(\hat{\mathbf{c}}_j^{(k)})}, \tag{11}$$

where $\mathbf{c}_j^{(i)}$ is the value of the i-th dimension of one-hot encoded community vector of c_j. The loss of self-supervised task in the region view O_r is defined in the same way.

Methods	BEIJING						SHANGHAI					
	RMSE			MAE			RMSE			MAE		
	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=1$	$\tau=2$	$\tau=3$
HA	17.75	20.40	24.55	8.84	10.30	12.34	21.39	21.39	21.39	9.02	9.02	9.02
LR	14.08	17.55	21.07	7.87	9.41	11.14	16.14	20.93	25.78	7.95	9.57	11.46
GBDT	13.52	16.02	18.25	7.64	8.70	9.68	14.38	16.59	18.57	7.19	7.94	8.68
GRU	14.09	17.24	19.59	7.98	9.33	10.35	15.86	19.31	21.96	7.76	8.89	9.87
STGCN	13.51	15.68	17.08	7.31	8.50	9.30	20.57	24.73	28.83	8.48	9.67	11.15
CoST-Net	14.67	15.62	16.31	7.87	8.38	8.73	17.08	20.26	21.19	7.76	8.63	9.04
ST-ResNet	13.88	14.48	16.17	7.27	7.60	8.31	15.30	16.42	18.73	7.08	7.83	8.59
Ada-MSTNet	**11.67**	**12.36**	**13.17**	**6.73**	**7.02**	**7.40**	**13.00**	**14.40**	**15.27**	**6.64**	**7.03**	**7.29**

Table 1: Overall performance of Ada-MSTNet and baselines given by RMSE and MAE on BEIJING and SHANGHAI.

By considering the MAE loss and self-supervised losses, our model aims to jointly minimize the following objective,

$$O = O_m + \alpha(O_r + O_c), \tag{12}$$

where α is the hyper-parameter controls the importance of two self-supervised losses.

Experiments

Experimental Setup

Data description. We use two real-world datasets, Beijing and Shanghai. All sensitive attributes such as user ID and phone number are anonymized for privacy concern. Both datasets are ranged from June 19, 2019 to July 26, 2019, and contains 25 communities. All communities are generated based on 7 attributes, *i.e.*, *Age*, *Gender*, *Life stage*, *Car owner*, *Education level*, *Income level* and *Consumption level*, which are either provided by user or mined from data. By considering both inflow and outflow demand, there are 50 tasks in total. Both datasets contain 784 adjacent $1\ Km \times 1\ Km$ regions in the city center. In particular, the Beijing dataset contains $125,042,989$ queries issued by $9,776,290$ users, whereas the Shanghai dataset contains $125,423,662$ queries issued by $9,311,549$ users. We chronologically order each dataset, set one hour as the unit time step, use the first 60% as the training set, the next 20% as the validation set, and the last 20% for testing.

Evaluation metrics. We use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), two widely used metrics (Zhang, Zheng, and Qi 2017; Ye et al. 2019; Zhang et al. 2020; Zhou and Tung 2015) for evaluation. To avoid overlook minorities, the reported overall performance is combined by all communities through averaged aggregation.

Baselines. We compare our model with three statistical learning methods and four deep learning models.

• **Historical Average (HA)** predicts the transportation demand by averaging the demand in the same periods, *e.g.*, extracting all demand in 8:00-9:00 from all historical Mondays to predict the demand of 8:00-9:00 in next Monday.

• **Linear Regression (LR)** is a classical machine learning method for regression task. We concatenate previous T steps historical demands as the input and predict each community demand separately.

• **Gradient Boost Regression Tree (GBRT)** is a variant of boosting model which is widely used in many data mining tasks. We use the version in XGBoost (Chen and Guestrin 2016) and the input is same as LR.

• **GRU** (Cho et al. 2014) is an efficient variant of recurrent neural network. It captures the temporal dependency but cannot handle spatial correlation.

• **ST-ResNet** (Zhang, Zheng, and Qi 2017) is another deep learning approach for transportation demand prediction. It captures spatial correlation with a deep residual network, but also overlooks the community correlation.

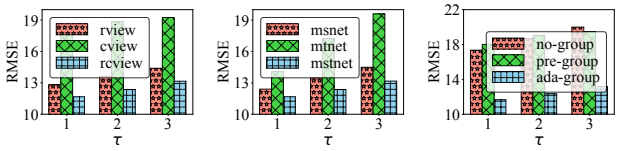
• **STGCN** (Yu, Yin, and Zhu 2018) is a graph neural network based model for traffic forecasting. It jointly models spatial and temporal correlation, but overlooks the community correlation.

• **CoST-Net** (Ye et al. 2019) is a co-prediction method via a LSTM based auto-encoder. We modify CoST-Net to simultaneously predict demands of multiple communities, by analogizing communities as different transport tools.

Implementation details. For fair comparison, all hyper-parameters of all baselines are carefully tuned. Besides, we train separate statistical learning models for different communities and train unified deep learning models for all communities by regarding the transportation demand of different communities as a multi-variate time series, since such formulation achieves a relatively better performance. We optimize the models by Adam (Kingma and Ba 2014). Specifically, we set the learning rate to 0.0001, hidden dimension $d = 512$ and $\alpha = 0.5$. We set input length $L = 18$ and $\tau = 3$ for prediction. The number of task groups is set to 4. The activation function in GNN is LeakyReLU with slope ratio 0.1. Follow exiting traffic prediction works (Yu, Yin, and Zhu 2018; Li et al. 2018), we apply Z-score normalization. All models run on a Linux server with Intel Xeon 5117 CPU, 128 GB Memory, and NVIDIA Tesla P40 GPU.

Overall Performance

Table 1 reports the overall performance of our model as well as all baselines on BEIJING and SHANGHAI with respect to RMSE and MAE. As can be seen, Ada-MSTNet consistently outperforms all baselines in terms of both metrics when $\tau = 1$, $\tau = 2$ and $\tau = 3$. Specifically, Ada-MSTNet outperforms all baselines at least (15.77%, 17.15%, 22.78%) and (8.02%, 8.26%, 12.3%) on BEIJING for (1 hour, 2 hour, 3 hour) prediction. The improvements of Ada-MSTNet on SHANGHAI are (10.62%, 14.03%, 21.61%) and (6.63%, 11.38%, 17.83%), respectively. Generally, the improvements of our model on further time steps are larger,



(a) Effect of multi-view graphs. (b) Effect of spatio-temporal network. (c) Effect of task grouping.

Figure 3: Ablation study of Ada-MSTNet given by RMSE.

demonstrating the advantage of Ada-MSTNet on distant future predictions. Looking further into the results of deep-learning baselines, we observe ST-ResNet achieves the state-of-the-art forecasting performance, which demonstrate its effectiveness on modeling spatial dependencies via a deep residual network. In addition, although CoST-Net jointly predicts all communities, we observe its performance is slightly worse than ST-ResNet. One possible reason is that it simply shares a network architecture for all different communities but does not model the nonuniform correlation among different communities; therefore, it introduces noises from less related community tasks. Lastly, we observe GBDT, as the state-of-the-art statistical learning baseline, is competitive with most deep-learning based baselines, which validates the effectiveness of GBDT on handling non-linear regression tasks.

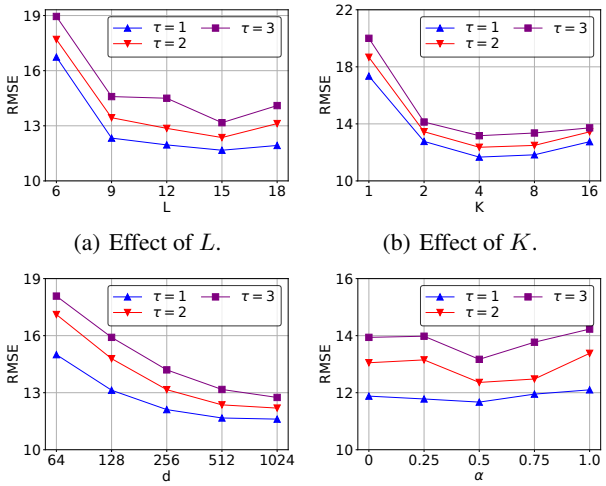
Ablation Study

We further look into Ada-MSTNet to verify the effectiveness of each component. Due to page limit, we only report the results on BEIJING by using RMSE, the results on BEIJING using MAE and on SHANGHAI are similar.

Effect of multi-view graphs. We first examine the effectiveness of multi-view graphs. Specifically, we evaluate three variants of Ada-MSTNet, (1) *rview* only includes region view graphs, (2) *cview* only involves community view graphs, and (3) *rcview* involves both graph views. As reported in Figure 3(a), Ada-MSTNet achieves the best performance by combing region view and community view, which demonstrates the benefit of constructing multi-view transportation demand graphs. Moreover, the RMSE of *rview* is lower than *cview*, indicating that the spatial dependency plays a more important role in CATDP.

Effect of spatio-temporal modeling. For spatio-temporal neural network, we evaluate the following variants, (1) *msnet* only involves GAT component, (2) *mtnet* only involves GRU component, and (3) *mstnet* combines both GAT and GRU component. As illustrated in Figure 3(b), jointly applying GNN and RNN results in the best prediction performance. We notice solely involving the GAT component achieves better performance than solely involving the GRU component. One possible reason is that we attach historical demands as features with each vertex in *msnet*, which includes temporal information to some extent.

Effect of adaptive task grouping. For the adaptive task grouping, we compare (1) *no-group* doesn't apply task grouping, (2) *pre-group* statically decides task grouping based on PCC as defined in Equation (4), and (3) *ada-group* apply adaptive task grouping. As shown in Figure 3(c), *no-group* or *pre-group* cannot share useful knowledge for CATDP, which further validates the necessity of adaptively clustered multi-task learning. Besides, we observe *no-group* performs slightly better than *pre-group*, which perhaps because of the inaccurate task grouping in *pre-group* introduced noises to less related tasks.



(a) Effect of L. (b) Effect of K.

(c) Effect of d. (d) Effect of α.

Figure 4: Parameter sensitivity on BEIJING.

Parameter Sensitivity

Then we investigate the parameter sensitivity of Ada-MSTNet. We report the impact of input length L, the number of task groups K, hidden dimension d, and task weight α on BEIJING using RMSE. When we vary a parameter, we keep the other parameters fixed as their default values.

First, we vary L from 6 to 18. The results are reported in Figure 4(a). Ada-MSTNet achieves lowest errors when $L = 15$. This illustrates a longer input sequence can provide more information to improve the prediction accuracy, but the information in distant previous steps are less useful.

Then, we vary K from 1 to 16. The results are illustrated in Figure 4(b). Ada-MSTNet achieves state-of-the-art performance when $K \geq 4$. Besides, we observe even clustering tasks into two groups is significantly helpful to reduce the prediction error. We choose $K = 4$ in the overall evaluation.

After that, we vary d from 64 to 1024. The results are demonstrated in Figure 4(c). By increasing d from 64 to 512, we observe a performance improvement. However, the prediction error becomes relatively stable when we further increase d to 1024. For efficiency and memory concern, we choose $d = 512$ in the overall evaluation.

Finally, to check the impact of self-supervised tasks, we vary α from 0 to 1. The results are reported in Figure 4(d). As can be seen, the influence of α is relatively small compared with the rest parameters. When we increase α, the prediction error of Ada-MSTNet first decreases then increases. Ada-MSTNet performs best when $\alpha = 0.5$. The above results illustrate that a moderately larger α can notably help cluster tasks into proper groups.

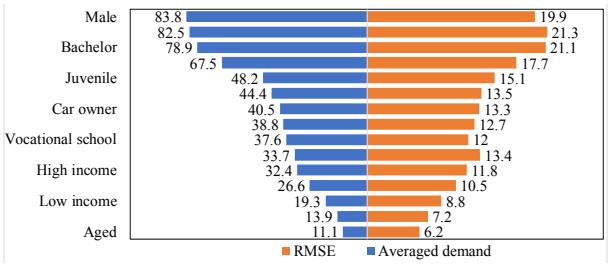


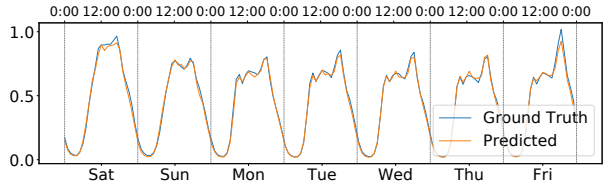
Figure 5: RMSE on representative communities.

Performance on Different Communities

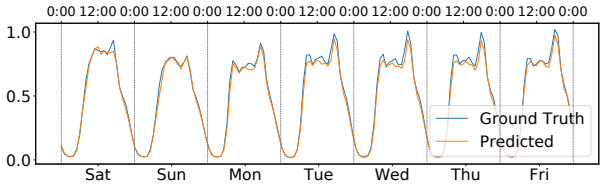
To evaluate the robustness of Ada-MSTNet, we further look into the prediction error in different communities. Figure 5 reports the prediction error and averaged transportation demand of fifteen representative communities. Overall, we observe the prediction error of different communities is correlated with the overall demand. One major reason perhaps is for the same ratio of demand fluctuates, communities with higher demand will result in higher RMSE. The above results suggest the future effort can be applied to these high transportation demand regions and communities to improve the overall performance.

Case Study

Finally, we qualitatively analyze the performance of Ada-MSTNet. Figure 6 illustrates the normalized transportation demand as well as predicted transportation demand of juveniles and the aged in one week from July 20, 2019 to July 26, 2019. Overall, our model successfully models daily periodicity and accurately predicted transportation demand. Impressively, our model accurately predicts the outlier peak of juveniles on Friday evening peak hours and the daily evening peak hour for the aged. The above results demonstrate that Ada-MSTNet effectively captures the characteristics of different communities and can be used for downstream tasks such as public resource allocation and community-aware advertisement.



(a) The observed and predicted demands of juveniles.



(b) The observed and predicted demands of the aged.

Figure 6: Qualitative study of prediction results.

Related Works

Traffic prediction. Previous studies on transportation demand or flow prediction are focus on the demands of the crowd collectively. For example, (Zhang, Zheng, and Qi 2017) adopted the convolution neural network for citywide region flow prediction. Based on emerging graph neural network techniques, (Li et al. 2018) and (Yu, Yin, and Zhu 2018) constructed road network graphs and apply GNN to model spatial dependency among road segments for better prediction. Moreover, (Yao et al. 2018) proposed a demand prediction model by learning dependencies on multi-graphs such as POI semantic correlation (Zhou et al. 2019). (Gu et al. 2020) proposed an interpretable framework for bike flow prediction. Recently, (Ye et al. 2019) investigated taxi and shared bike demand joint prediction by using a collective recurrent auto-encoder. However, those methods are designed for collective transportation demand or traffic flow prediction, none of them investigates the diversified transportation demand of different communities.

Multi-task learning. Multi-task learning (MTL) is a learning paradigm that leverages shared information in related tasks to improve model generalization ability (Zhang and Yang 2017). Recently, multi-task learning has been widely adopted for spatio-temporal problems. For instance, (Zheng and Ni 2013) proposed a multi-task regression framework to capture the temporal dynamics of travel cost. (Deng et al. 2017) leveraged the generalization capability of MTL to make predictions under different traffic situations. (Liu et al. 2021) proposed a hierarchical MTL model to derive unified route representation learning for multi-modal transportation recommendation (Liu et al. 2020). The above works simply combine a few highly related tasks, but neglect the nonuniform correlation among a large number of tasks. In recent studies, (Kang, Grauman, and Sha 2011; Zhou and Zhao 2015; Yao, Cao, and Chen 2019) proposed soft clustering strategies to improve task performance. However, theses methods either require pre-computed group assignment or only work on linear models. In this work, we leverage the supervision signal from two different views to achieve end-to-end soft task grouping.

Conclusion

In this paper, we proposed Ada-MSTNet, a soft clustered multi-task spatio-temporal neural network for city-wide community-aware transportation demand prediction (CATDP). Specifically, we first constructed a sequence of multi-view transportation demand graphs to characterize the time-evolving spatial adjacency and community demand similarity. After that, we devised a spatio-temporal network for simultaneous spatio-temporal and community correlation modeling. By regarding the transportation demand prediction of each community in different regions as individual tasks, we proposed an adaptively clustered multi-task learning module to selectively share common knowledge among highly related tasks. In particular, a mutually supervised task grouping strategy is proposed to decide task group assignment by leveraging supervision signals from the one another graph view in an end-to-end fashion. Finally, extensive experimental results demonstrated the effectiveness of Ada-MSTNet on two real-world large-scale datasets.

References

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, 1–4.

Chen, T.; and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734.

Deng, D.; Shahabi, C.; Demiryurek, U.; and Zhu, L. 2017. Situation aware multi-task learning for traffic prediction. In *IEEE International Conference on Data Mining*, 81–90.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Gu, J.; Zhou, Q.; Yang, J.; Liu, Y.; Zhuang, F.; Zhao, Y.; and Xiong, H. 2020. Exploiting Interpretable Patterns for Flow Prediction in Dockless Bike Sharing Systems. *IEEE Transactions on Knowledge and Data Engineering* .

Kang, Z.; Grauman, K.; and Sha, F. 2011. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 521–528.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations*.

Liu, H.; Han, J.; Fu, Y.; Zhou, J.; Lu, X.; and Xiong, H. 2021. Multi-Modal Transportation Recommendation with Unified Route Representation Learning. *Proceedings of the VLDB Endowment* 14(3): 342–350.

Liu, H.; Tong, Y.; Han, J.; Zhang, P.; Lu, X.; and Xiong, H. 2020. Incorporating Multi-Source Urban Data for Personalized and Context-Aware Multi-Modal Transportation Recommendation. *IEEE Transactions on Knowledge and Data Engineering* .

Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; and Damas, L. 2013. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems* 14(3): 1393–1402.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5998–6008.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph attention networks. In *6th International Conference on Learning Representations*.

Yao, H.; Wu, F.; Ke, J.; Tang, X.; Jia, Y.; Lu, S.; Gong, P.; Ye, J.; and Li, Z. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2588–2595.

Yao, Y.; Cao, J.; and Chen, H. 2019. Robust task grouping with representative tasks for clustered multi-task learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1408–1417.

Ye, J.; Sun, L.; Du, B.; Fu, Y.; Tong, X.; and Xiong, H. 2019. Co-prediction of multiple transportation demands based on deep spatio-temporal neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 305–313.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-semporal graph convolutional networks: A deep learning framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3634–3640.

Zhang, J.; Zheng, Y.; and Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 1655–1661.

Zhang, W.; Liu, H.; Liu, Y.; Zhou, J.; and Xiong, H. 2020. Semi-supervised hierarchical recurrent graph neural network for city-wide parking availability prediction. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 1186–1193.

Zhang, Y.; and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* .

Zheng, J.; and Ni, L. M. 2013. Time-dependent trajectory regression on road networks via multi-task learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 1048–1055.

Zhou, J.; Gou, S.; Hu, R.; Zhang, D.; Xu, J.; Jiang, A.; Li, Y.; and Xiong, H. 2019. A collaborative learning framework to tag refinement for points of interest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1752–1761.

Zhou, J.; and Tung, A. K. 2015. Smiler: A semi-lazy time series prediction system for sensors. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1871–1886.

Zhou, Q.; and Zhao, Q. 2015. Flexible clustered multi-task learning by learning representative tasks. *Transactions on Pattern Analysis and Machine Intelligence* 38(2): 266–278.