Learning Geo-Contextual Embeddings for Commuting Flow Prediction

Zhicheng Liu[1,2], **Fabio Miranda**[2], **Weiting Xiong**[3,4], **Junyan Yang**[3], **Qiao Wang**[1], **Claudio T. Silva**[2]

[1]School of Information Science and Engineering, Southeast University, Nanjing, China
[2]Tandon School of Engineering, New York University, New York, USA
[3]School of Architecture, Southeast University, Nanjing, China
[4]Media Lab, Massachusetts Institute of Technology, Boston, USA
{zhichengliu, qiaowang}@seu.edu.cn, {fmiranda, csilva}@nyu.edu, weitingx@mit.edu, yjy-2@163.com

Abstract

Predicting commuting flows based on infrastructure and land-use information is critical for urban planning and public policy development. However, it is a challenging task given the complex patterns of commuting flows. Conventional models, such as gravity model, are mainly derived from physics principles and limited by their predictive power in real-world scenarios where many factors need to be considered. Meanwhile, most existing machine learning-based methods ignore the spatial correlations and fail to model the influence of nearby regions. To address these issues, we propose **G**eo-contextual **M**ultitask **E**mbedding **L**earner (**GMEL**), a model that captures the spatial correlations from geographic contextual information for commuting flow prediction. Specifically, we first construct a geo-adjacency network containing the geographic contextual information. Then, an attention mechanism is proposed based on the framework of graph attention network (GAT) to capture the spatial correlations and encode geographic contextual information to embedding space. Two separate GATs are used to model supply and demand characteristics. A multitask learning framework is used to introduce stronger restrictions and enhance the effectiveness of the embedding representation. Finally, a gradient boosting machine is trained based on the learned embeddings to predict commuting flows. We evaluate our model using real-world datasets from New York City and the experimental results demonstrate the effectiveness of our proposal against the state of the art.

Introduction

The commute of people from home to work is a phenomenon that has shaped society and cities throughout the ages, from ancient Egypt to modern New York City (Bram and McKay 2005; Austin 2017). These daily recurrent movements form a complex network that is highly correlated with the socioeconomic factors of cities (Simini et al. 2012; Spadon et al. 2019). Access to public services, open spaces, transportation and even entertainment all play a role that influence where a worker will live, or where a company will station its offices.

In order to have more efficiently planned cities, it is crucial to understand how commuting flows are impacted by
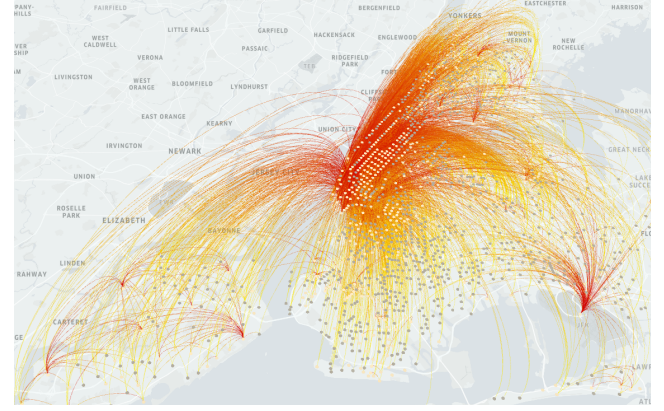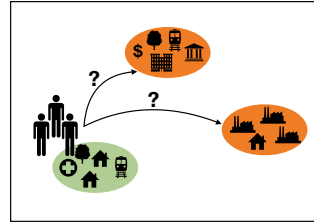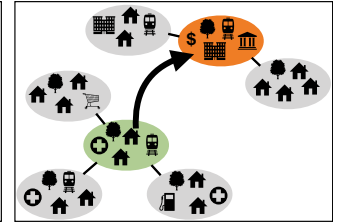
Figure 1: Overview. (a) Example of the commuting flow network of New York City in 2015. Yellow indicates origin census geographic units, and red indicates destination units. (b) Illustration of the commuting flow prediction problem when residents choose where to work based on supporting infrastructure, distance, etc. (c) Illustration of our solution that uses geographic contextual information for commuting flow prediction.

infrastructure and land use. This information can be used in urban planning to guide the development of new districts, or in transportation planning to direct the deployment of new modes of transport. Consider for instance the example of Manhattan, a dense borough of New York City. Because of its historical concentration of transit hubs and major corporations, it is the principal destination for workers

from the outer boroughs of the city (see Fig. 1a). Redevelopment and rezoning initiatives, however, have contributed to the increase in the number of jobs located in Brooklyn and Queens (Office of the State Deputy Comptroller for the City of New York 2019). Understanding the commuting flow can then help answer many what-if questions in the planning stage, such as *"If a new high-tech industrial park is planned for a region in Brooklyn, from which regions would people commute to work? How should we plan the supporting infrastructure to improve the commuting efficiency?"*, which could help urban planners, policy makers and different stakeholders to make informed decisions.

As such, commuting flow prediction is one of the fundamental problems for urban planning in that it reveals the spatial interactions of supply and demand in a city (Rodrigue, Comtois, and Slack 2016). The problem differs from spatiotemporal traffic origin-destination (OD) forecasting problem. Traffic OD forecasting is essentially a time series prediction problem where the historical movements will be used as input features, while commuting flow prediction problem aims at revealing spatial interaction of supply and demand in a city by predicting the edge-level signals (e.g. the volume of the flow), using **only** the information of node attributes, such as infrastructure and land use information (see Fig. 1b). Since commuting behaviour shows a daily repeated static pattern, this property enables us to develop a connection between commute and urban indicators.

Although the problem has a long history that goes back to the eighteenth century (Monge 1781), it remains challenging because of the inherent complexity of cities. Past proposals that use gravity or radiation model (Lenormand, Bassolas, and Ramasco 2016; Simini et al. 2012) make simple assumptions of the generation process of commuting flows, and might not capture certain commute patterns. Recently proposed machine learning-based models, such as gradient boosting machine (Pourebrahim et al. 2019), on the other hand, only consider the features of origin-destination, ignoring the influence of nearby regions.

To address the above issues, we propose a model called Geo-Contextual Multitask Embedding Learner (GMEL) to predict commuting flows. First, we construct a geographic adjacency network that represents the geographic relationship between regions (Fig. 1c). Then, inspired by Tobler's first law of geography (Tobler 1970) which states that *everything is related to everything else, but nearby things are more related than distant things*, we employ Graph Attention Network (GAT) to encode the geographic contextual information into an embedding space. Since commuting flow can be viewed as a kind of spatial interaction between origin supply and destination demand (Rodrigue, Comtois, and Slack 2016), we employ two separate GATs to encode the geographic contextual information into two different embedding spaces. This allows us to disentangle the supply and demand characteristics that are hidden in the infrastructure and land use features. To ensure the effectiveness of the embeddings representation, multitask objective functions is used to introduce stronger restrictions forcing the embeddings to encapsulate effective representation for flow prediction (Caruana 1997). Finally, we train a gradient boosting machine re-

gression model using the embeddings generated by GMEL as input features.

The primary contributions of this paper are the following:

- We propose the use of geographic contextual information for commuting flow prediction problem. To our knowledge, we are the first to exploit geographic contextual information in this task.

- We propose a model (GMEL) to capture the spatial correlations from geographic contextual information and encode the information into embedding space based on graph attention network.

- We conduct extensive experiments using real-world datasets from New York City. The results demonstrate the effectiveness of our proposed method against the state of the art.

Related Work

Commuting Flow Prediction

In this paper, we focus on commuting flow prediction problem (Spadon et al. 2019). It is important to note that the problem formulation is different from the traffic origin-destination (OD) forecasting problem (Xiong et al. ; Iwata and Hitoshi 2019). Although both of the problems view human movements as networks, traffic OD forecasting problem is essentially an edge-level time series prediction problem where the historical network edges (e.g. ODs) can be the input features of the model (Wang et al. 2019b), while commuting flow prediction problem aims at predicting the edge weights (e.g. the volume of the flow) utilizing only the attribute of nodes.

Gravity model (Lenormand, Bassolas, and Ramasco 2016) is a widely used conventional model for commuting flow prediction which makes simple assumptions of the generation process of commuting flow. For example, the original form of gravity model (Zipf 1946) assumes the number of commuters traveling from one region to another is proportional to the product of the population of origin and destination and decays with the distance of the trip, as shown below:

$$\hat{T}_{ij} = \alpha_0 \frac{M_i^{\alpha_1} M_j^{\alpha_2}}{d_{ij}^{\alpha_3}} \tag{1}$$

where M_i and M_j are the population of region i and j respectively, d_{ij} is the travel distance between the two regions, and $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ are the parameters to be estimated. However, this simple assumption might not capture the complex nature of a city (Albeverio et al. 2007).

Recently, researchers have used nonparametric models, such as gradient boosting machine, to capture the complex nature of spatial interaction represented by commuting flow (Pourebrahim et al. 2019; Robinson and Dilkina 2018). These off-the-shelf machine learning models usually present better performance than conventional physics-derived models. However, these models simply use origin-destination node attributes as input features to fit a regression model, ignoring the influence of nearby regions. Another family of conventional models is called intervening

opportunity model. These models consider the influence of nearby potential competitors of origin or destination, such as radiation model (Simini et al. 2012; Yang et al. 2014). Inspired by the idea of intervening opportunity, we propose using the geographic contextual information to develop the regression model where the embeddings of each node is encoded with the influence of nearby regions.

Graph Representation Learning

Several graph representation learning methods have been proposed recently. A general inductive framework called GraphSAGE is proposed by (Hamilton, Ying, and Leskovec 2017), which leverages node attribute to generate node embeddings in a message-passing way. Also, graph attention network (Veličković et al. 2018) leverages self-attention mechanism to allow messages passed by neighbors to be aggregated with different weights. Motivated by these works, we use the framework of graph attention network and adapt the attention mechanism to our tasks so that our model could capture the geographic context.

Several applications have also been proposed based on graph neural network. (Pan et al. 2019) utilized a GAT-like structure to learn embeddings that capture spatial correlations of traffic patterns. (Wang et al. 2019b) proposed a GraphSAGE-like graph embedding model to capture the spatial mobility patterns and neighboring correlations. (Zhang et al. 2019) proposed a multitask learning framework to simultaneously predict the node and edge traffic flows. Few studies explored the use of graph neural network to capture spatial correlations for commuting flow predictions.

Preliminaries

In this section, we introduce the definitions and problem formulation.

Definition 1 **Urban Geographic Unit**: We partition the city into N urban geographic units $v_1, v_2, ..., v_N$. The geographic units can be street blocks, census tracts, zip code areas, etc.

Definition 2 **Urban Indicators**: The urban indicator $\mathbf{a}_i$ is a vector that serves as the attribute of v_i. It characterizes the aggregated information of infrastructure and land use of the geographic units.

Definition 3 **Geo-Adjacency Network**: The Geo-adjacency network is an undirected weighted graph $G_{adj} = (V, E, A)$ where $V = \{v_1, v_2, ..., v_N\}$ is the set of urban geographic units which serves as the nodes of the graph, $E = \{e_{ij}|v_i, v_j \text{ are geographically adjacent}, 1 \leq i, j \leq N\}$ is the set of edge features that describes the strength of correlations (e.g. travel distance, trip duration) and $A = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$ is the set of urban indicators that serves as the node attributes.

In our case, we use census tracts as the urban geographic unit. The geo-adjacency network of New York City is shown in Fig. 2.

Definition 4 **Distance Matrix**: The distance matrix D is a N-by-N matrix where the entry D_{ij} represents the travel distance from v_i to v_j.

Definition 5 **Commuting Trips**: Commuting trips are a set of triplets $T = \{(v_i, v_j, T_{ij})\}$ where v_i is the trip origin
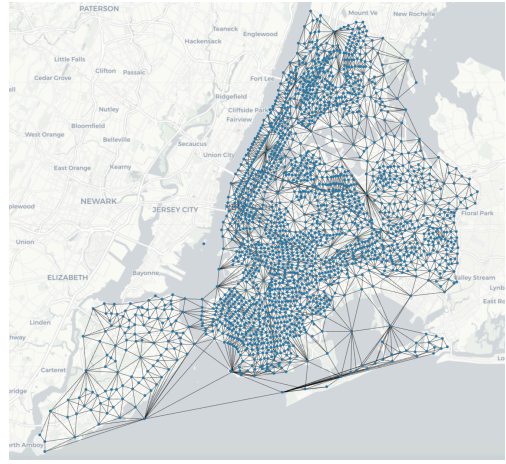

Figure 2: Geo-adjacency network of New York City. The dots represent the centroids of census tracts and the lines represent the edges.

node, v_j is the trip destination node, T_{ij} is the commuting flow, i.e. the number of commuters travel from v_i to v_j. Note that T_{ij} can be seen as an edge-level flow. We also define two kinds of node-level flows, i.e. in-flow and out-flow. We denote the out-flow as $T_{i:}$ representing the total number of outgoing commuters from v_i and denote the in-flow as $T_{:j}$ representing the total number of incoming commuter to v_j.

Problem: Given $G_{adj} = (V, E, A)$ and D, develop a regression model to predict $T_{ij} \in T$.

Methodologies

In this section, we describe the architecture of our model for commuting flow prediction. Basically, our model consists of two components: Geo-contextual Multitask Embedding Learner and Flow Predictor.

1) **Geo-contextual Multitask Embedding Learner (GMEL)**. GMEL is designed to capture the spatial correlations from geographic context. Basically, the geographic context can be viewed as the graph neighborhoods of G_{adj}. GMEL utilizes Graph Attention Network (GAT) to encode the geographic contextual information into an embedding space. To disentangle the supply and demand characteristics that are hidden in infrastructure and land use, GMEL employs two separate GATs to encode the geographic contextual information into two different embedding space. To ensure the effectiveness of the embeddings representation, GMEL employs multitask learning framework which imposes stronger restrictions forcing the embeddings to encapsulate effective representation for flow prediction (Caruana 1997).

2) **Flow Predictor**. Considering the learned embeddings from GMEL, we employ gradient boosting machine (GBM) as the regression model to predict commuting flows. GBM has advantages in handling dense numerical features (Ke et al. 2019), such as the learned embeddings in our scenarios. By iteratively evaluating the largest information gain of features, GBM can automatically select and combine use-
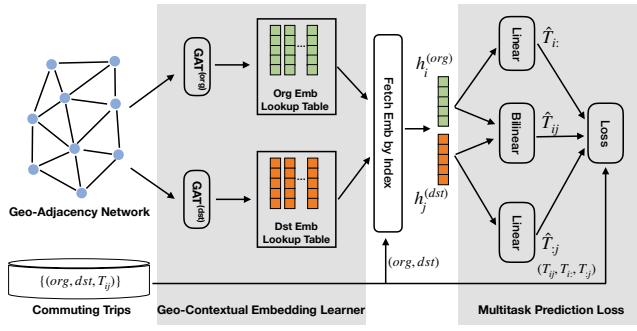
Figure 3: Framework of GMEL

ful numerical features to fit the targets (Friedman, Hastie, and Tibshirani 2001). This is why most recently proposed machine learning models for commuting flow prediction employ gradient boosting regression tree (GBRT) or random forest as the regression function (Spadon et al. 2019; Pourebrahim et al. 2019; Robinson and Dilkina 2018). In particular, we use GBRT in this paper.

Framework

The framework of GMEL is shown in Fig. 3. GMEL aims at learning effective embeddings of urban geographic units which encode the geographic contextual information. To learn the supply and demand characteristics for each geographic unit respectively, we employ two separate GATs to encode this information. The generated embeddings are then applied to a bilinear function to predict the flow. Meanwhile, these embeddings will also be applied to two linear functions to predict the in/out-flow of the geographic units. The overall prediction loss is the weighted sum of the three tasks' loss, and we use backpropagation to train GMEL in an end-to-end manner.

Graph Attention Network

Graph attention network (GAT) iteratively aggregates the information from node neighborhood and updates the node states with nonlinearity. The weight used to aggregate the neighborhood messages depends on the features of two connecting nodes and edge features.

Assume the state of node i is $h_i^{(l)} \in \mathbb{R}^{m \times 1}$ in the l-th layer and the features of edge (v_i, v_j) is $e_{ij} \in \mathbb{R}^{n \times 1}$. GAT first applies linear transformation to these vectors.

$$z_i^{(l)} = W^{(l)} h_i^{(l)} \tag{2}$$

$$c_{ij}^{(l)} = V^{(l)} e_{ij} \tag{3}$$

where $W^{(l)} \in \mathbb{R}^{k \times m}$ and $V^{(l)} \in \mathbb{R}^{t \times n}$ are trainable parameter matrices. The resulting z_i is the message vector passed to neighbors. Before aggregating these message vectors, an attention score for each edge is calculated:

$$r_{ij}^{(l)} = \sigma(a^{(l)T}(z_i^{(l)} || c_{ij}^{(l)} || z_j^{(l)})) \tag{4}$$

where $\sigma(\cdot)$ is a nonlinear function (e.g. Relu, Sigmoid), $a^{(l)} \in \mathbb{R}^{(2k+t) \times 1}$ is a trainable parameter vector that maps

the concatenation of messages into a scalar value and $||$ denotes the concatenation operation. Then, the attention scores are normalized by softmax:

$$\alpha_{ij}^{(l)} = \frac{exp(r_{ij}^{(l)})}{\sum_{k \in \mathcal{N}(i)} exp(r_{ik}^{(l)})} \tag{5}$$

where $\mathcal{N}(i)$ denotes the graph neighborhood of the i-th node. The final aggregation process consists of two parts representing the neighborhood impact and self impact respectively:

$$h_i^{(l+1)} = \sigma(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} z_j^{(l)} + U^{(l)} h_i^{(l)}) \tag{6}$$

where $U^{(l)} \in \mathbb{R}^{k \times m}$ is a trainable parameter matrix and $\mathcal{N}(i)$ is the neighborhood of node i.

Modeling Supply and Demand Characteristic

Commuting flows can be viewed as a kind of spatial interactions between supplies and demands (Rodrigue, Comtois, and Slack 2016). Our model holds an underlying assumption that the flows are determined by the supply characteristic of the origin geographic unit and the demand characteristic of the destination geographic unit.

To model both supply and demand characteristic of each geographic unit, we use two separate GATs. In Fig. 3, the GAT$^{(org)}$ extracts demand characteristic from origin geographic units and encodes the characteristic into origin embeddings. The GAT$^{(dst)}$ extracts supply characteristic from destination geographic units and encodes the characteristic into destination embeddings. The two GATs have the same structure, but the attention mechanism in GAT will assign different weights to different features based on the origin or destination roles, thus modeling supply and demand characteristics.

Multitask Learning

As the goal of GMEL is to learn embeddings that encode supply and demand characteristic for commuting flow prediction, we adopt multitask learning framework to put more restrictions for the GMEL training process.

Main Task: Predicting Commuting Flow Having the origin and destination embeddings $h_i^{(org)}$ and $h_j^{(dst)}$ from GATs, a bilinear model is used to predict the commuting flow:

$$\hat{T}_{ij} = h_i^{(org)T} W_b h_j^{(dst)} \tag{7}$$

where $W_b \in \mathbb{R}^{m \times m}$ is a trainable parameter matrix modeling the interactions between origin embeddings and destination embeddings. The corresponding loss function of the main task is:

$$\mathcal{L}_{main} = \frac{1}{|T|} \sum_{i,j} (\hat{T}_{ij} - T_{ij})^2 \tag{8}$$

where $|T|$ is the total number of trips.

Subtasks: Predicting In/Out Flow   We include prediction of the in/out-flow as two subtasks, i.e. predicting the total number of incoming/outgoing commuters of each geographic unit. The intuition is that the commuting flows and in/out-flows are highly correlated and, thus, the two subtasks would impose stronger restrictions on the training process of GMEL. The in/out-flows are predicted by linear functions:

$$\hat{T}_{i:} = \mathbf{w}_{out}^T h_i^{(org)} \tag{9}$$

$$\hat{T}_{:j} = \mathbf{w}_{in}^T h_j^{(dst)} \tag{10}$$

where $\mathbf{w}_{out}, \mathbf{w}_{in} \in \mathbb{R}^{m \times 1}$ are trainable vector parameters. The corresponding loss function of two subtasks are:

$$\mathcal{L}_{out} = \frac{1}{N}(\hat{T}_{i:} - T_{i:})^2 \tag{11}$$

$$\mathcal{L}_{in} = \frac{1}{N}(\hat{T}_{:j} - \hat{T}_{:j})^2 \tag{12}$$

where N is the number of geographic units.

Overall Loss Function   The overall loss function of GMEL is formulated as the weighted sum of all three tasks:

$$\mathcal{L}_{GMEL} = \lambda_{main}\mathcal{L}_{main} + \frac{\lambda_{sub}}{2}(\mathcal{L}_{in} + \mathcal{L}_{out}) \tag{13}$$

where $\lambda_{main}, \lambda_{sub}$ are the hyperparameters representing the weights for main task and subtasks respectively.

Training Algorithm

Recall that our model consists of two components: GMEL and flow predictor. We train GMEL using stochastic gradient descent method in an end-to-end manner. The learning process of GMEL can be seen as pre-training. Having the embeddings from well-trained GMEL, a GBRT is trained as flow predictor based on the concatenation of origin-destination embeddings and travel distance to predict the commuting flow. The training process is summarized in Algorithm 1.

Experiments

In this section, we provide an empirical evaluation of our proposed model on real-world dataset.

Datasets

We validate our proposed model on real-world datasets from New York City. To make comparison with state-of-the-art models in the literature, we use similar experimental settings as reported in (Pourebrahim et al. 2019). We use the 2010 New York City census tracts as geographic units (2168 units in total). For commuting trips and urban indicators, we use the following datasets:

LODES   The 2015 Origin-Destination Employment Statistics (LODES) dataset presents the commuting trips of interest (US Census Bureau 2015). It is collected yearly and records the home and employment locations of workers, representing stable commuting flows. These flows are aggregated into geographic unit level flow. 3,031,641 commuters and 905,837 pairs of origin-destination trips were

---

Algorithm 1: Training Algorithm

Input: Geo-adjacency Network $G_{adj} = (V, E, A)$,
Distance Matrix D,
Commuting Trips $T_{train} = \{(v_i, v_j, T_{ij})\}$

Output: The learned GMEL,
The learned flow predictor $\hat{f}$

1 /* GMEL Learning   */
2 **repeat**
3 │ $T_{batch} \leftarrow$ Draw a training batch from T_{train}
4 │ $\{h_i^{(org)}\} \leftarrow GAT^{(org)}(G_{adj})$
5 │ $\{h_j^{(dst)}\} \leftarrow GAT^{(dst)}(G_{adj})$
6 │   Evaluate $\mathcal{L}_{GMEL}$ by $(\{h_i^{(org)}\}, \{h_j^{(dst)}\}, T_{batch})$
   │   using Equation 13
7 │ $\nabla\mathcal{L}_{GMEL} \leftarrow$ Backpropagate $\mathcal{L}_{GMEL}$
8 │ $w \leftarrow w - \gamma\nabla\mathcal{L}_{GMEL}$ // γ is the
   │   learning rate
9 **until** *stopping criterion is met*;

10 /* Flow Predictor Learning   */
11 $\{h_i^{(org)}\} \leftarrow GAT^{(org)}(G_{adj})$
12 $\{h_j^{(dst)}\} \leftarrow GAT^{(dst)}(G_{adj})$
13 $\mathcal{X}_{input} \leftarrow \{\}, \mathcal{Y}_{input} \leftarrow \{\}$
14 **for** (v_i, v_j, T_{ij}) *in* T_{train} **do**
15 │ $\mathcal{X}_{input} \leftarrow \mathcal{X}_{input} \cup Concat(h_i^{(org)}, h_j^{(dst)}, D_{ij})$
16 │ $\mathcal{Y}_{input} \leftarrow \mathcal{Y}_{input} \cup T_{ij}$
17 **end**

18 $\hat{f} \leftarrow$ Train GBRT on $(\mathcal{X}_{input}, \mathcal{Y}_{input})$

---

collected in New York City. We randomly divide the commuting trips into training, validation and test datasets by 6:2:2.

PLUTO   The 2015 NYC Primary Land Use Tax Lot Output (PLUTO) presents the urban indicators of interest (NYC DCP 2015). It records land use and infrastructure information at the tax lot level. This information is aggregated into census tract level (65 urban indicators for each census tract). A summary of the urban indicators is listed in Table 1.

OSRM   We employ Open Source Routing Machine (OSRM) to measure the travel distance between the centroids of census tracts (Luxen and Vetter 2011). The travel distances will serve as the edge features of the geo-adjacency network.

Baselines

To show the effectiveness of our model, we compare our model with the following baselines:

- **Gravity Model with Power-Law Decay (GM-P)**: Gravity model with power-law distance decay function is the most classic model for spatial interaction model. It's widely used in predicting commuting flows, cargo shipping volume, etc. Basically, gravity model is a log-linear model. The difference of models in this family lies in the

Table 1: Summary of Urban Indicators

Categories	# Features	Contents
Infrastructure	40	The number of different types of buildings (25), the density of commercial/residential/etc. units (4), the number of buildings in each built year interval (11)
Land Use	23	The number of tax lots in different land use (11), the land area ratio of retail/office/etc. (10), statistics of floor area ratio (2)
Speciality	2	Whether or not the census tract contains landmarks or historic districts (2)
Total	65	

form of the distance decay function. For further details of gravity model, we refer the readers to (Lenormand, Bassolas, and Ramasco 2016).

- **Gravity Model with Exponential Decay (GM-E)**: Gravity model with exponential distance decay function is another model in gravity model family. It is reported to have better performance in predicting commuting flows (Lenormand, Bassolas, and Ramasco 2016).

- **Random Forest (RF)**: Recently, researchers proposed to use gradient boosting machine to predict commuting flows. RF is reported as the state-of-the-art model (Pourebrahim et al. 2019; Spadon et al. 2019).

- **Gradient Boosting Regression Tree (GBRT)**: GBRT belongs to gradient boosting machine family and is widely used as regression model.

- **Node2vec**: Node2vec is an unsupervised learning model to learn node embeddings from graph structured data (Grover and Leskovec 2016). Recently, its variant have also been applied to learn location embeddings for spatio-temporal prediction tasks (Wang and Li 2017). We incorporate Node2vec to learn the embeddings for each census tract on the geo-adjacency network and use these embeddings as inputs to train a gradient boosting regression tree as flow predictor.

To validate the effectiveness of our model architecture, we also implement two variants of our model:

- **GMEL-noMul**: We remove multitask settings and only keep the main task, i.e. setting $\lambda_{main} = 1, \lambda_{sub} = 0$ in Equation 13.

- **GMEL-noSep**: We remove the settings of using two separate GATs to model supply and demand characteristics respectively. Instead, only one GAT is used to generate embeddings and this set is used as both origin and destination embeddings.

The above baseline models are compared to the GMEL with multitask weights to be $\lambda_{main} = 0.5, \lambda_{sub} = 0.5$, the embeddings size of both GATs to be 128, and number of GAT layers to be 2.

We implemented our model and the baselines using PyTorch (Paszke et al. 2017) and Deep Graph Library (Wang et al. 2019a). The experiments were executed on a Intel E5-2690 v4 2.6 GHz, 256 GB of RAM, and a NVIDIA Tesla P100 GPU with 12 GB of RAM.

Evaluation Metrics

To measure the prediction performance, we adopt three evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Common Part of Commuters (CPC).

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{i,j} (\hat{T}_{ij} - T_{ij})^2} \quad (14)$$

$$MAE = \frac{1}{|T|} \sum_{i,j} |\hat{T}_{ij} - T_{ij}| \quad (15)$$

$$CPC = \frac{2 \sum_{ij} min(\hat{T}_{ij}, T_{ij})}{\sum_{ij} \hat{T}_{ij} + \sum_{ij} T_{ij}} \quad (16)$$

RMSE and MAE are widely used as evaluation metrics for regression problem. CPC is widely used in commuting flow prediction problem (Lenormand, Bassolas, and Ramasco 2016; Robinson and Dilkina 2018), and it measures the common part of agreements between predicted value and target value. CPC is 0, when no agreement is found, and is 1, when the two are identical.

Performance Analysis

Table 2: Performance on Test Set

Model	RMSE	MAE	CPC*
GM-P	7.035	2.236	0.589
GM-E	6.944	2.179	0.602
RF	6.273	2.436	0.638
GBRT	5.454	1.974	0.707
Node2vec	5.455	1.994	0.704
GMEL-noMul	5.356	1.910	0.716
GMEL-noSep	4.982	1.772	0.737
GMEL (ours)	**4.887**	**1.747**	**0.741**

* Higher is better.

We evaluate the performance of the baseline models and our model on the test set, and summarize the results in Table 2. From the experiments, we have the following observations:

(a) Effect of number of GAT layers. (b) Effect of embedding size. (c) Effect of multitask weights.
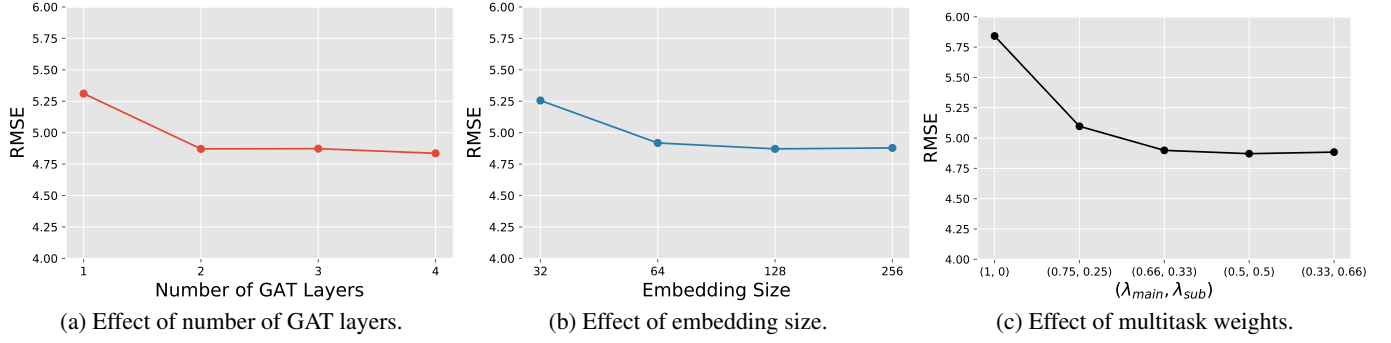
Figure 4: Results of different hyperparameter settings.

- Gravity models have the worst performance among all models. The reason might be that simple assumptions of gravity model cannot capture the complex patterns of commuting flows and thus lead to poor predictive power.

- RF and GBRT generally have better performance than gravity models, which is in accordance with recently published literature (Pourebrahim et al. 2019). The reason might be that gradient boosting machine is better capable of handling nonlinearity.

- Node2vec is slightly better than RF and only comparable to GBRT, even though it uses graph neighborhood structure to generate embeddings. The reason might be Node2vec is designed to preserve network neighborhood of nodes, but this neighborhood information is not useful for characterizing the supply and demand characteristics of the city.

- All GMEL variants outperform the above baseline models. This verifies the effectiveness of leveraging geographic contextual information for commuting flow prediction.

- GMEL outperforms GMEL-noMul and GMEL-noSep. This shows the effectiveness of multitask learning framework and the necessity of modeling supply and demand characteristic separately.

To this end, we have validated the effectiveness of our model.

Residual Analysis

To illustrate the effectiveness of exploiting spatial correlations, we present the residual maps in Fig. 5. These maps show the difference between predicted and ground-truth incoming flows, i.e. the sum of the residuals of flows to the same destination, in each census tract. We compare GMEL with the state-of-the-art model GBRT (Pourebrahim et al. 2019). In Fig. 5, we can observe that the residuals of GMEL are spatially smoother than that of GBRT. The reason is that GMEL exploits geographic contextual information to capture spatial correlations, and in doing so the prediction will take into account both the characteristics of regions of interest and the influence of nearby regions.
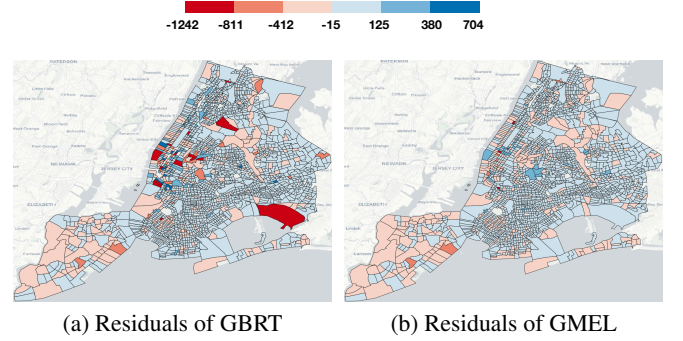

(a) Residuals of GBRT (b) Residuals of GMEL

Figure 5: Spatial distribution of residuals. Red indicates underestimation and blue indicates overestimation. Light blue census tracts indicate the best predictions.

Parameter Sensitivity Analysis

We also analyze the parameter sensitivity of our model. Three main hyperparameters of GMEL are examined, namely the number of GAT layers, embedding size and multitask weights. The results are shown in Fig. 4.

The effect of number of GAT layers The number of GAT layers determines the depth of graph neighborhood. For example, if the number of GAT layers is 2, then all graph neighboring nodes within two-hops would have an effect on the target node. In our scenario, this hyperparameter implicitly defines the geographic range of influence. From Fig. 4a, we can see that when the number of GAT layers is one, the model performs worse. When the number of GAT layers is greater than or equal to two, the performance doesn't fluctuate too much. This indicates that the effective graph neighborhood is two-hops graph neighborhood. In New York City, the two-hops graph neighborhood covers on average 1.5 km, which is approximately 15-minutes walking distance.

The effect of embedding size We also conduct experiments on several alternatives of embedding size, i.e. 32, 64, 128, 256. In Fig. 4b, we can find that the performance increases as the embedding size increases from 32 and saturates at the size of 128.
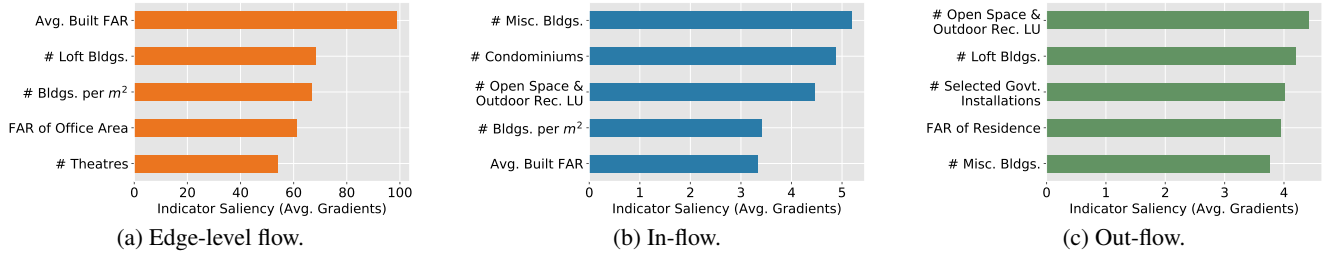
Figure 6: Top-5 salient urban indicators.

The effect of multitask weights Different set of multitask weights are also tested, see Fig. 4c. Recall that the subtasks are introduced to enhance the performance of the main task. Indeed, we can observe in Fig. 4c that when the weight of subtasks increases, the performance of the main task keeps increasing until the weights of the main task and subtasks are equal, i.e. $\lambda_{main} = 0.5, \lambda_{sub} = 0.5$.

Feature Sensitivity Analysis

We evaluate the impact of the urban indicators by computing the saliency map of GMEL (Simonyan, Vedaldi, and Zisserman 2013). In our case, saliency map represents the average gradients of output estimates with regards to urban indicators, exhibiting its overall effect on commuting flows. A larger absolute value of the saliency map points to a more prominent urban indicator. Three saliency maps are evaluated: edge-level flows, in-flows and out-flows. Fig. 6 shows the most prominent urban indicators. These salient urban indicators present the supply and demand characteristics for different kind of flows. For example, the number of buildings per square meters, indicating job opportunities, is salient for in-flow, meanwhile, floor area ratio of residence, indicating the density of regular residences, is salient for out-flow.

Case Study

To further evaluate the usefulness of our proposal, we show a case study focusing on census tracts that experienced major changes in their urban indicators between the years of 2013 and 2015. We first select a set of 5 census tracts that had the largest changes when considering their urban indicators between these two years. Next, we train a model on the 2013 data set (PLUTO and LODES) and test the prediction performance of the model considering the 2015 data set. In our experiments, the mean absolute error and standard deviation between the predicted flow values and the groundtruth for the selected 5 census tracts are the following: $1.07 \pm 1.25, 1.40 \pm 2.12, 1.26 \pm 2.08, 0.99 \pm 1.73, 0.86 \pm 1.37$. By having a model trained on a particular year, GMEL can be used to predict the origin and destination of new commuting flows, given changes in the urban indicators. This highlights how our proposal can guide urban planners and policy makers to make informed decisions when it comes to new urban development scenarios.

Conclusion

In this paper, we study the problem of predicting commuting flow using only the information of infrastructure and land use, a fundamental problem in urban planning and public policy development. Different from conventional gravity model and recently proposed machine learning methods, we propose the use of geographic contextual information for commuting flow prediction. As such, an end-to-end embedding learning framework based on graph attention network is proposed to learn geo-contextual embeddings of the geographic units. The learned embeddings are then fed to a gradient boosting machine to make predictions. We conduct extensive experiments on real-world datasets from New York City. The results show that introducing geographic contextual information can greatly improve the accuracy of prediction and our model outperforms all baseline methods including the state of the art.

Acknowledgements

References

Albeverio, S.; Andrey, D.; Giordano, P.; and Vancheri, A. 2007. *The dynamics of complex urban systems: An interdisciplinary approach.* Springer.

Austin, A. E. 2017. The cost of a commute: A multidisciplinary approach to osteoarthritis in new kingdom egypt. *International Journal of Osteoarchaeology* 27(4):537–550.

Bram, J., and McKay, A. 2005. Evolution of commuting patterns in the New York City metro area. *Current Issues in Economics and Finance*.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*, volume 1. Springer Series in Statistics, New York, NY, USA.

Grover, A., and Leskovec, J. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 855–864. New York, NY, USA: ACM.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 1024–1034.

Iwata, T., and Hitoshi, S. 2019. Neural collective graphical models for estimating spatio-temporal population flow from aggregated data. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:3935–3942.

Ke, G.; Xu, Z.; Zhang, J.; Bian, J.; and Liu, T.-Y. 2019. DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 384–394. New York, NY, USA: ACM.

Lenormand, M.; Bassolas, A.; and Ramasco, J. J. 2016. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography* 51:158–169.

Luxen, D., and Vetter, C. 2011. Real-time routing with open-streetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, 513–516. New York, NY, USA: ACM.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

NYC DCP. 2015. Pluto and mappluto. https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.

Office of the State Deputy Comptroller for the City of New York. 2019. New York City employment trends.

Pan, Z.; Liang, Y.; Wang, W.; Yu, Y.; Zheng, Y.; and Zhang, J. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 1720–1730. New York, NY, USA: ACM.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NeurIPS Workshop*.

Pourebrahim, N.; Sultana, S.; Niakanlahiji, A.; and Thill, J.-C. 2019. Trip distribution modeling with Twitter data. *Computers, Environment and Urban Systems* 77:101354.

Robinson, C., and Dilkina, B. 2018. A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, 30:1–30:8. New York, NY, USA: ACM.

Rodrigue, J.-P.; Comtois, C.; and Slack, B. 2016. *The geography of transport systems*. Routledge.

Simini, F.; González, M. C.; Maritan, A.; and Barabási, A.-L. 2012. A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Spadon, G.; Carvalho, A. C. P. L. F. d.; Rodrigues-Jr, J. F.; and Alves, L. G. A. 2019. Reconstructing commuters network using machine learning and urban indicators. *Scientific Reports* 9(1):1–13.

Tobler, W. 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography* 46(2):234–240.

US Census Bureau. 2015. Longitudinal employer-household dynamics. https://lehd.ces.census.gov/data/.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Wang, H., and Li, Z. 2017. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 237–246. New York, NY, USA: ACM.

Wang, M.; Yu, L.; Zheng, D.; Gan, Q.; Gai, Y.; Ye, Z.; Li, M.; Zhou, J.; Huang, Q.; Ma, C.; Huang, Z.; Guo, Q.; Zhang, H.; Lin, H.; Zhao, J.; Li, J.; Smola, A. J.; and Zhang, Z. 2019a. Deep graph library: Towards efficient and scalable deep learning on graphs. *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Wang, Y.; Yin, H.; Chen, H.; Wo, T.; Xu, J.; and Zheng, K. 2019b. Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 1227–1235. New York, NY, USA: ACM.

Xiong, X.; Ozbay, K.; Jin, L.; and Feng, C. Dynamic origin-destination matrix prediction with line graph neural networks and Kalman filter. 23.

Yang, Y.; Herrera, C.; Eagle, N.; and González, M. C. 2014. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports* 4:5662.

Zhang, J.; Zheng, Y.; Sun, J.; and Qi, D. 2019. Flow prediction in spatio-temporal networks based on multitask deep learning. *IEEE Transactions on Knowledge and Data Engineering* 1–1.

Zipf, G. K. 1946. The P1 P2/D hypothesis: On the intercity movement of persons. *American Sociological Review* 11(6):677–686.