CD-CNN: A Partially Supervised Cross-Domain Deep Learning Model for Urban Resident Recognition

Jingyuan Wang, Xu He, Ze Wang, Junjie Wu, Nicholas Jing Yuan, Xing Xie, Zhang Xiong
Beihang University
Microsoft Corporation, Microsoft Research
Research Institute of Beihang University in Shenzhen

Introduction

The CD-CNN model is concerned with three core problems in mobile data-driven resident recognition:

- **How to extract users' behavioral features from mobile phone data.**

We decompose the mobile phone signaling records into two domains: **the location domain and the communication domain**. Convolutional neural networks are adopted to extract behavioral features from high-dimensional raw data of both domains.

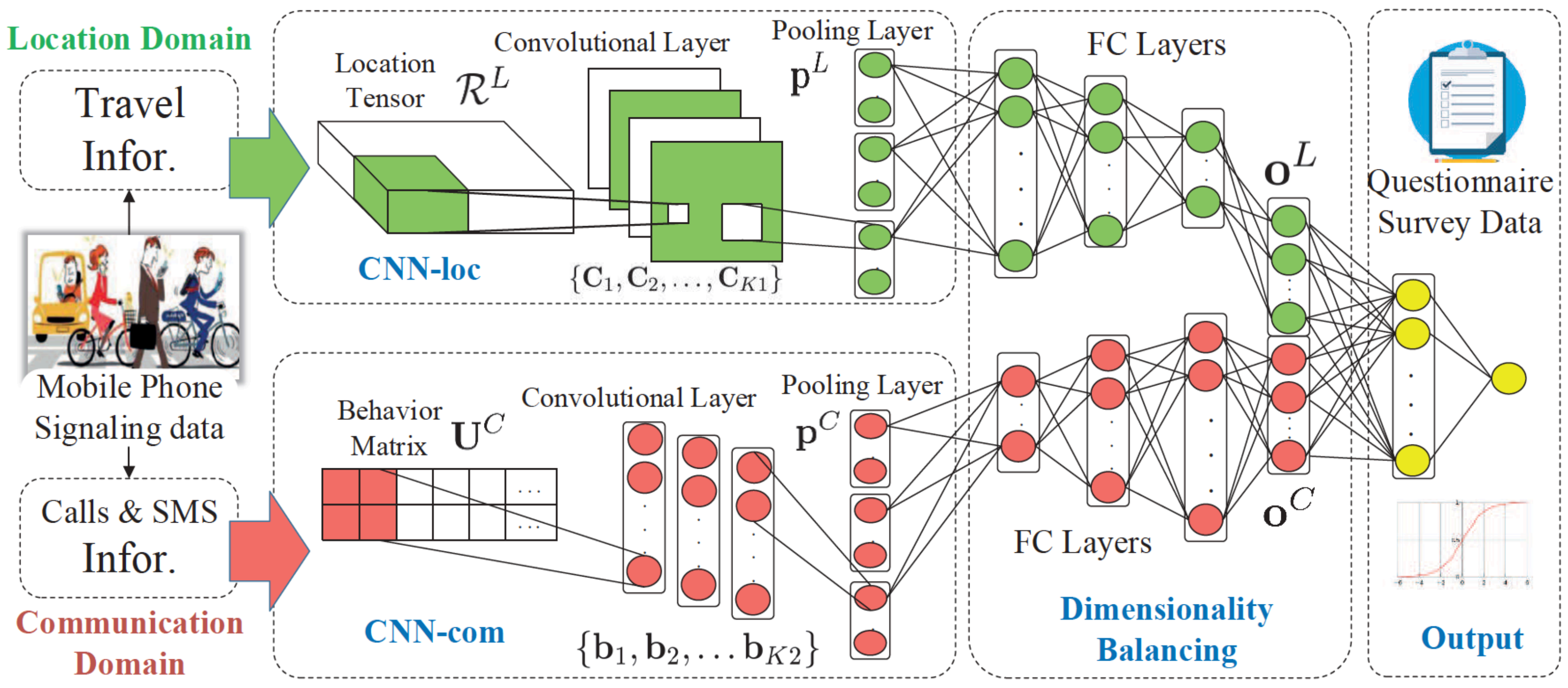
- **How to fuse knowledge from multiple domains.**

For heterogeneous and severely imbalanced features generated by CNNs in the location and communication domains, respectively, we introduce a carefully designed **dimensionality balancing mechanism** for knowledge fusion, which is crucial for the success of classification.

- **How to handle incomplete label information of data sets.**

In our study, only a very small part of mobile phone users are labeled by volunteer questionnaire surveys. To deal with this, we plug **a co-training scheme** into the pretraining/fine-turning framework of deep learning, which solves the cross-domain learning and partially supervised learning problems simultaneously.

The CD-CNN Model



CNN for Communication Domain
We divided the areas of a city into ***I*J*** square zones, and divide a day into 24 time slices. We define location matrices R^w and R^h for a resident to represent the normalized total hours of staying in each zone in working period and home period.
And we use a neural network consists of a convolutional layer and a pooling layer to extract feature from the location matrics.

$$
\begin{aligned}
c_{pq} \;&=\; \sigma\Bigg(u + \sum_{m=1}^{M}\sum_{n=1}^{N} a^h_{nm} r^h_{p+n,q+m} \\
&\qquad + \sum_{m=1}^{M}\sum_{n=1}^{N} a^w_{mn} r^w_{p+m,q+n}\Bigg)
\end{aligned}
$$

CNN for Communication Domain
From the mobile phone signaling data, we extract two types of operation behavior information, i.e. , calls and short messages. For a mobile phone user in the time slice t , we calculate e_t and s_t as the number of the calls and short messages normalized by total call and SMS volume of a user, respectively. A communication matrix of a resident is expressed as

$$
\mathbf{U}^C = \begin{bmatrix} e_1, & e_2, & \dots, & e_t, & \dots, & e_{24} \\ s_1, & s_2, & \dots, & s_t, & \dots, & s_{24} \end{bmatrix}
$$

The convolution layer generates convolution neuron vectors as:

$$
b_n = \sigma\left(v + \sum_{h=1}^{H} a^e_h e_{n+h} + \sum_{h=1}^{H} a^s_h s_{n+h}\right)
$$

Cross-domain Co-training for CD-CNN

In many real-life cases, labeling a sample is very costly. For example, in our Wuxi case, we have only 30 thousands well-labeled data. But we have a large scale of unlabeledmodel data.

In order to exploit the information in both labeled and unlabeled data, we propose a partially supervised network training algorithm based on the co-training scheme, named as Cross-domain Network Co-training (CNC).
The CNC algorithm contains three training steps:

- Domain separated pretraining

CNC uses labeled samples to respectively train the LN model and the CN model to obtain the first step optimized parameters.

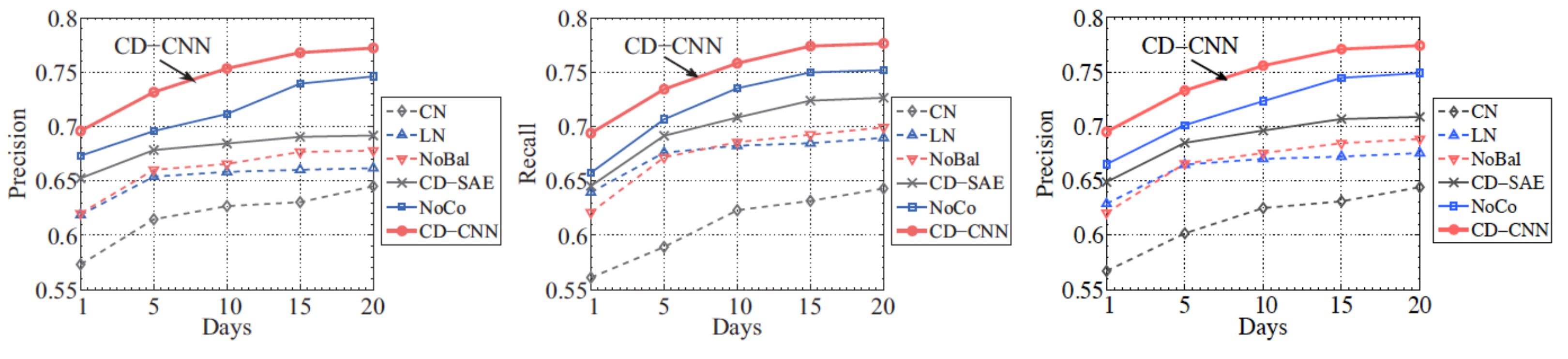
- Domain crossed co-training

The CNC algorithm uses unlabeled samples to collate the LN and CN models each other in an iterative way.

- Supervised finetuning

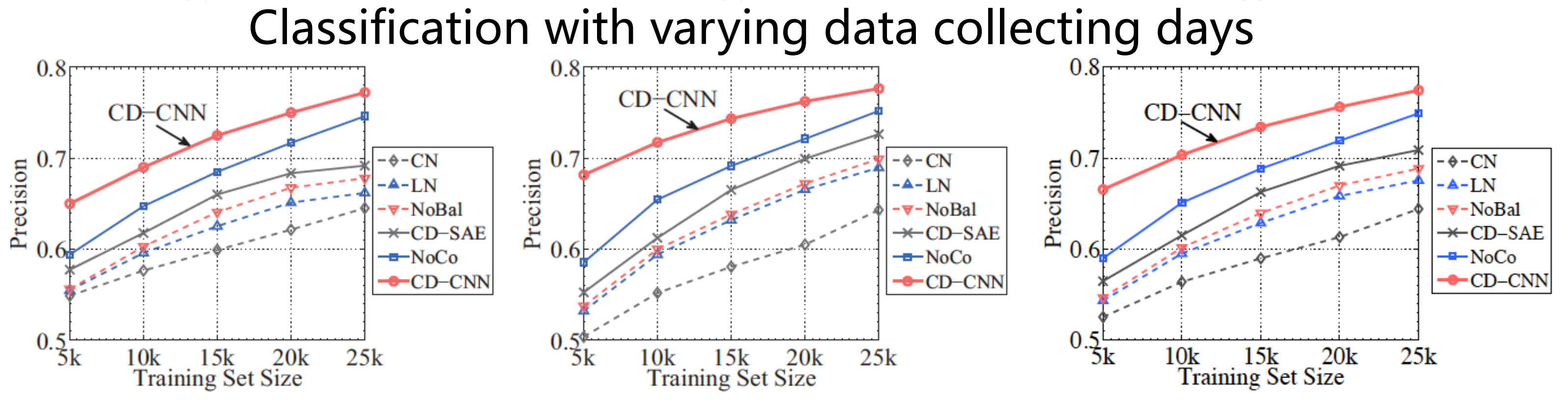
In the fine-tuning step, the CNC algorithm once again uses labeled samples to train the model

Experiment Results



(a) Precision (b) Recall (c) F1 Score

Classification with varying data collecting days



(a) Precision (b) Recall (c) F1 Score

Classification with varying sizes of labeled samples

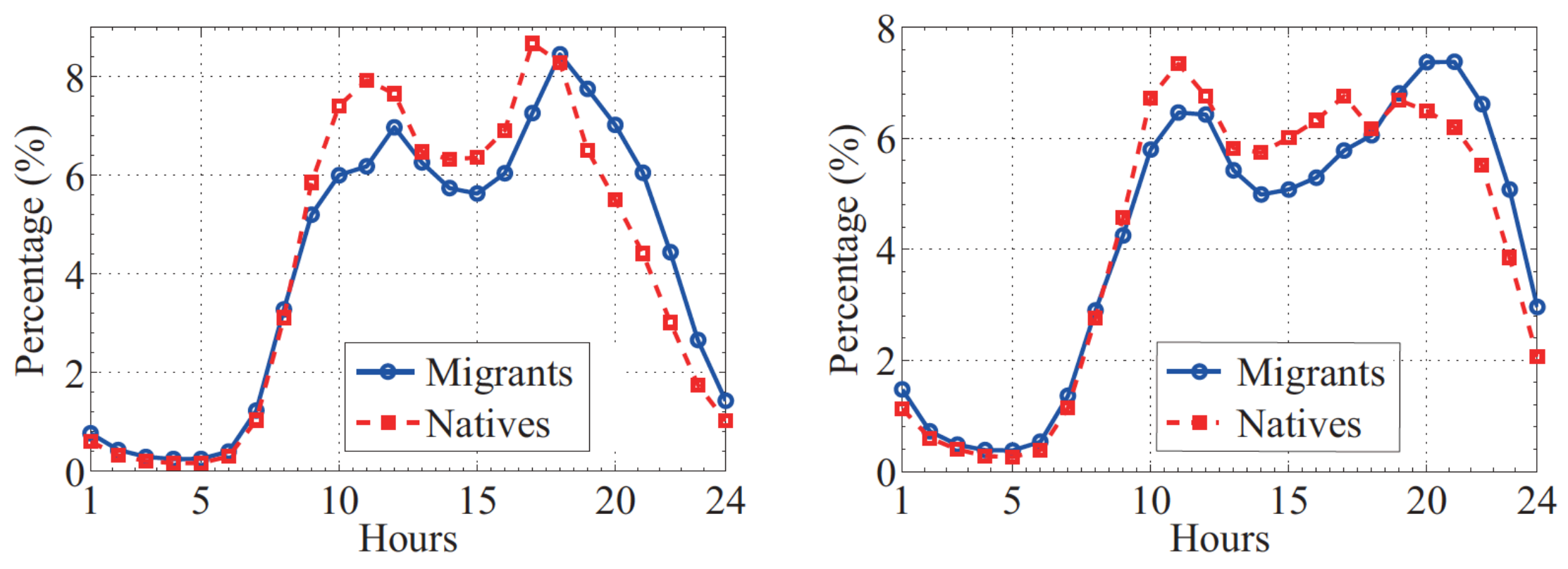
Applications

- Population Census

Using the CD-CNN model, we precisely predicted the population census. The result shows the 35% residents in Wuxi are migrants and the remaining 65% are natives, which properly matched with the data of the Statistical Yearbook of Wuxi.

- Temporal distribution comparison of call and message

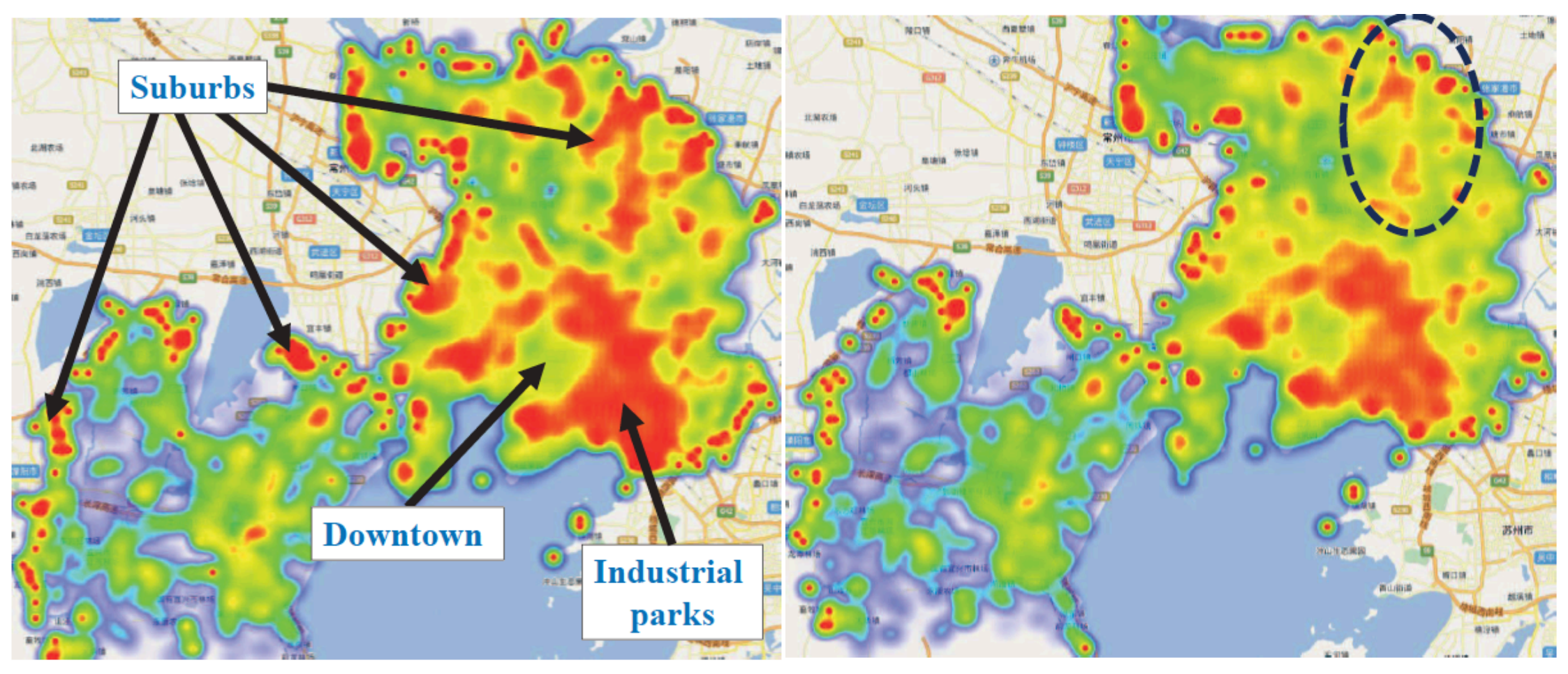
The evening peak of calls and short messages for migrants is later than the natives.



(a) Calls (b) SMS

- Migrant distributions during the home and working periods

The color of the map expresses the proportion of migrants to total residents in an area — the redder, the higher. As shown in the map, two types of areas have higher migrant proportions:
i) the areas surrounding the downtown, especially in the industrial parks;
ii) the suburbs of the city.



(a) Home Period (b) Working Period