Incorporating Location Based Social Networks in the Prediction of Real-Time Taxi Demand
with Deep Learning

Dong He, Yang Chen
School of Computer Science, Fudan University, China
{dhe15, chenyang}@fudan.edu.cn

INTRODUCTION

Problem:

- Real-time taxi demand prediction is important in a smart city.
- Existing approaches has minor attention to external information, and thus fail to achieve very accurate predictions.

Proposal:

- Integrating the taxi data with around 1 million user check-in records collected from a LBSN, the Swarm App.
- Applying a Phased LTSM network for the prediction model.

PREDICTION MODEL

Preliminaries:

- **Taxi demand**: the number of taxi pick-ups in a region during a time step, denoted as y_t^n.
- **Check-in number**: the number of user check-ins from the venues in a region during a time step, denoted as g_t^n
- **Taxi demand prediction**: to predict the taxi demand $\widehat{y}_{t+1}^n$ at time step i_{t+1}, given the taxi demand and check-in number before time step i_t, as well as the spatial-temporal features.

Proposed Approach:

- Our objective is to find the most accurate $\mathcal{F}$ in the following equation,

$$\widehat{y}_{t+1}^n = \mathcal{F}(y_{t-L,\dots,t}^R, g_{t-L,\dots,t}^R, \boldsymbol{q}_{t-L,\dots,t}^R)$$

  R: the set of all regions, n: the n-th region,
  $y_{t-L,\dots,t}^R$: taxi demand in the previous L time steps,
  $g_{t-L,\dots,t}^R$: check-in numbers in the previous L time steps,
  $\boldsymbol{q}_{t-L,\dots,t}^R$: spatial-temporal features in the previous L time steps, and $\boldsymbol{q}_t^n = (weekday, latitude, longitude)$.
- Figure 1 shows the overview of our Phased LSTM-based approach. Figure 2 shows the cell architecture of our Phased LSTM network.

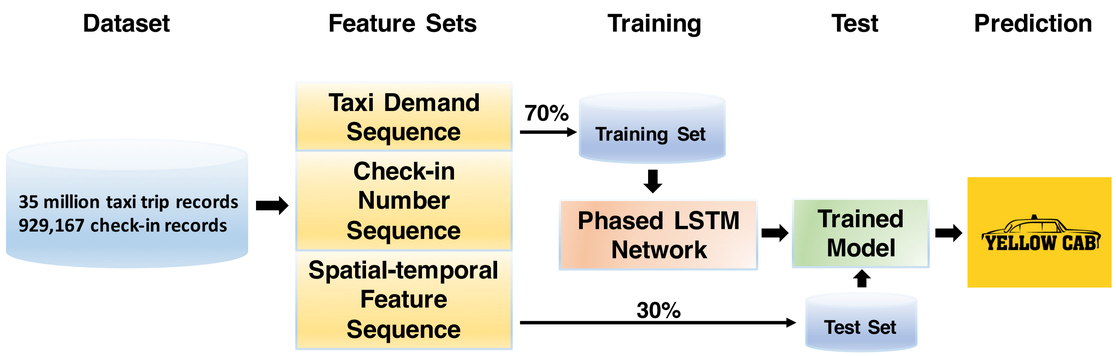

Figure 1: Overview of our Phased LSTM-based approach

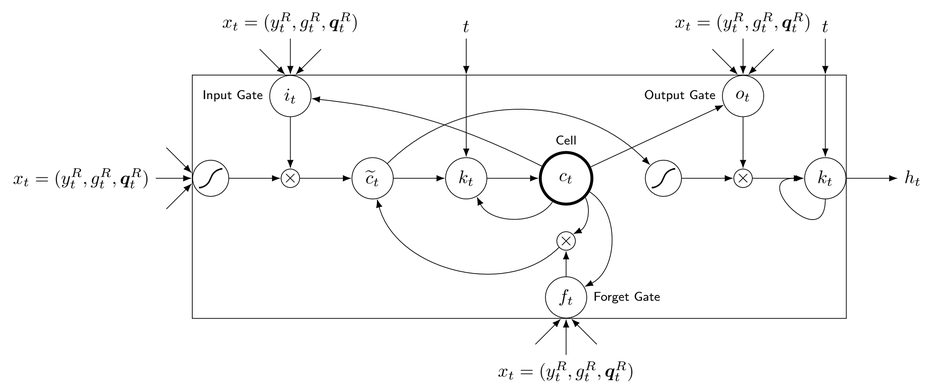

Figure 2: Architecture of the cell of our Phased LSTM network

- At time step i_t, the input to the network is $x_t = (y_t^R, g_t^R, \boldsymbol{q}_t^R)$. The output of the network is the prediction of the taxi demand in the region at time step i_{t+1}. We aggregate the taxi pick-ups and check-in numbers to get the taxi demand y_t^n and check-in number g_t^n, as well as the spatial-temporal features $\boldsymbol{q}_t^n$.

EXPERIMENTS

Dataset:

- Taxi trip data: 35 million taxi trip records in NYC.
- Swarm check-in data: 929,167 check-ins in NYC.

Metrics:

- Mean Absolute Percentage Error (MAPE).
- Root Mean Square Error (RMSE).

Results:

- As shown in Table 1, the Phased LSTM-based approach that incorporates LBSNs achieves the lowest MAPE (0.1547) and the lowest RMSE (12.03).
- Compared to the LSTM model which does not incorporate LBSNs, our approach has a relative improvement of 21.27% on MAPE and 6.96% on RMSE, which corroborates that LBSNs carry useful information to infer the taxi demand.
- Figure 3 shows that our standard LSTM-based and Phased LSTM-based approaches that incorporate LBSNs consistently outperform others on all seven days in a week.

Table 1: Comparisons with different baseline methods

Method	MAPE	RMSE
Linear Regression	0.2397	13.42
Random Forest Regression	0.2156	13.31
XGBoost Regression	0.2187	13.23
LSTM without external data	0.1965	13.24
Our standard LSTM-based model with external data	0.1658	12.30
Our Phased LSTM-based model with external data	**0.1547**	**12.03**

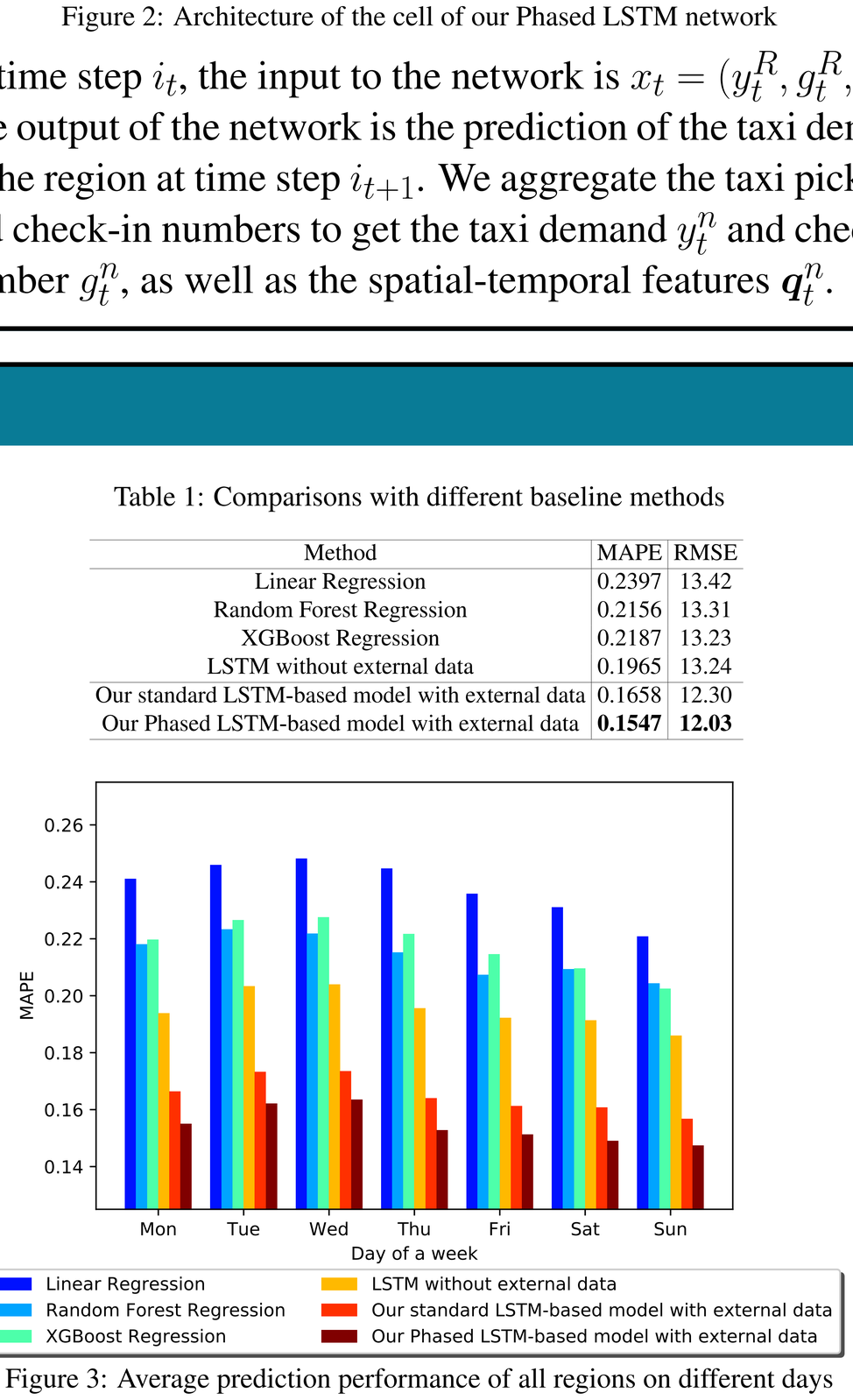

Figure 3: Average prediction performance of all regions on different days