数据驱动方法在城市中的应用

新问题、新思路与新方法

詹仙园 博士
京东智能城市研究院资深研究员

THU 2018/12/13
About Me

2007 - 2011
B.E.: Structure Engineering at THU

2011 - 2012
M.S.: Transportation Engineering at Purdue

2013 - 2017
Ph.D.: Transportation Engineering at Purdue

2014 - 2016
M.S.: Computer Science at Purdue

2017 - 2018
Microsoft Research Asia

2018 - Present
JD iCity
• 新问题
  – 新数据 & 新问题
• 新思路
  – 数据驱动思维
• 新方法
  – 时空深度学习 & More
新问题
新数据 & 新问题
Massive Amount of Data in Urban Space
 Massive self generated data from urban space

• Social media data
  – 3 million Foursquare check-ins per day (2011)
  – 500 million tweets per day on Twitter (2012)

• Mobile phone data
  – 30 million mobile phone records per day
    (Great Boston area, 2009)

• GPS data from taxis
  – 500,000 taxi trips from NYC (2013)

• Data from urban sensor networks
  – E.g. License-plate recognition camera data
    40,000 vehicle records per day for a single intersection
    (Langfang, 2015)
Tremendous Opportunities

• Game changer for urban systems modeling
• Allows us directly observe how system works
• Better solution for existing problems:
  – Traffic state monitoring
  – Inferring land use
• New possibilities for emerging problems:
  – Large-scale logistics/dispatching optimization
  – Water/air quality estimation/prediction
  – Detecting illegal parking
  – And many more
Why Data-driven Methods

• Urban systems are highly complex
  – Millions of residents, large number of interacting sub-components
  – Simple analytical model simply will not work

• Requirement for efficiency and scalability
  – Solving analytical models are costly
  – Not suitable nor accurate for real-time applications

• Usability
  – Traditional analytic sometimes hardly useful for solving real world problems
Urban services involve logistics/dispatching optimization:

- Huge volumes of requests
- Large amount of data
- Real-time operation
- Highly dynamic

Optimization matters!
Highly dynamic
- Requests changing over time and locations
- States of resources: locations, load, tasks,…

Large-scale
- Thousands of candidates: locations, taxis, carriers…
- Correlated (cannot be simply separated)
- Most problems are NP-hard (scale is a disaster)

Instantaneous answers

Multiple constraints: time, cost, capacity…

- Use real-world data to mine patterns
- Use real-time data in models

- New data-driven algorithms
- Problem/region partitioning
- Search space reduction

- Utilizing efficient data processing/management techniques, e.g. spatial-temporal indexing

- Highly customized models
- Candidate pruning
AI改進救護車站點選址和調度優化
基于共享单车数据的城市违章停车智能监测

城市中违章停车随时随处可见

T. He, J. Bao, R. Li, S. R., Y. Li, C. Tian, Y. Zheng. Detecting Vehicle Illegal Parking Events using Sharing Bikes' Trajectories. KDD 2018
新思路
数据驱动思维
Traditional engineering approach:

- Behavior/system assumptions → Analytical model → Less accurate, non-scalable results
  - Limited data for validation

Conventional data-driven approach:

- Many sparse, weak data features → Limited information → Data-driven model → Results with reasonable accuracy
Feature engineering integrating data property and domain knowledge:

Many sparse, weak data features

Information fusion
Strong feature
Rich information
Data-driven model
Accurate results

Highly customized data-driven models integrating domain knowledge:

Many sparse, weak data features

Limited information
Hybrid data-driven model
Accurate results

Domain knowledge, etc.
城市计算(Urban Computing)

城市数据的采集、管理、分析挖掘和服务提供
数据 + 计算
解决交通、规划、环境、能耗、公共安全、商业、医疗等痛点

云计算 + 大数据 + AI + 城市场景
新方法
时空深度学习 & More
Taxonomy of Spatio-Temporal (ST) Data

- Data Structures
- Spatio-temporal (ST) Properties

- Spatio-temporal Static Data
  - POI Distributions
    - Bing, Google, Gaode, Baidu Maps
  - Road/Transportation Networks
    - Gaode Maps, Traffic management Bureau

- Spatial Static Temporal Dynamic Data
  - Weather/AQI Station Data
  - Road Traffic Data

- Spatio-Temporal Dynamic Data
  - Foursquare, Geo-tweets, Dianping
  - Trajectory Data
    - TAXI, DD, Uber, China Mobile, China Telecom

- US EPA, China MEP, IOT
Why Spatio-Temporal Data Is Unique

Spatial Properties

- **Distance**
  - Spatial closeness
  - Triangle inequality:
    \[ |d_1 - d_2| \leq d_3 \leq |d_1 + d_2| \]

- **Hierarchy**
  - Different spatial granularities
  - City structures

Why Spatio-Temporal Data Is Unique

• Temporal properties
  – Temporal closeness
  – Period
  – Trend

A) Hourly traffic speed on consecutive days
B) Traffic speed at 9-10am on consecutive weekends
Deep Learning meets ST Data

• What Deep Learning can do for ST Data
  – Encoding a (single) ST dataset
  – Fusing multiple ST datasets

• What ST data can provide to Deep Learning
  – Massive and diverse Data
  – Computing infrastructures are ready
  – Application scenarios requiring
    • Instantaneous responses at large spaces
    • Collective computing
    • (traditional machine learning models many not be able to handle)
Taxi Trajectory Data of Shenzhen
Encoding Spatio-Temporal Properties

**CNN** is able to model **spatial** properties

![CNN Diagram](image)

**RNN/LSTM** is able to model **temporal** properties

![RNN/LSTM Diagram](image)

**Pyramid Architecture:** **Hierarchy**
Trajectories of taxis, trucks and buses
Fusing Multiple ST-Datasets
Why Deep Learning for ST Data

• Big ST-Data

Traditional ML algorithms cannot model spatial and temporal properties of such a live and large-scale data.
Challenges of DL for ST Data

- Data transformation
- Encoding ST properties in DNNs

A) Hourly traffic speed on consecutive days
B) Traffic speed at 9-10am on consecutive weekends
AI预测城市栅格区域人群流量
Challenges

• Urban crowd flow depends on many factors
  – Flows of previous time interval
  – Flows of nearby regions and distant regions
  – Weather, traffic control and events

• Capturing spatial properties
  – Spatial distance and hierarchy

• Capturing temporal properties
  – Temporal closeness
  – Period and trend
Converting Trajectories into Video-like Data

ST-ResNet: A Collective Prediction

Residual Deep Convolutional Neural Network

Capturing spatial correlation of both near and far

Using residual network framework to help training

AI预测城市区域人流量及流转
Multi-view Graph Convolutional Networks

Junbo Zhang et al. Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. Submitted
Yuxuan Liang, Songyu Ke Junbo Zhang et al. GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. IJCAI 2018
Geo-sensory Time Series

- There are massive sensors deployed in physical world

### Properties
- Each sensor has a unique geospatial location
- **Constantly** reporting time series readings about different measurements
- With **geospatial correlation** between their readings

- RC: 0.84
- pH: 7.1
- Turbidity: 0.54
- Volume: 32
- Speed: 50km/h
- PM2.5: 74
- PM10: 90
- NO2: 57
Challenges

• Dynamic inter-sensor correlations
• Dynamic temporal correlations
• Affected by many factors
  – Readings of previous time interval
  – Readings of other sensors in nearby regions
  – External factors: weather, time and land use
GeoMAN: Multi-level Attention Networks

- Spatial attention to capture complex spatial correlations
- Temporal attention to model dynamic temporal correlations
- Fusion module to incorporate the external factors

Spatial Attention

- Local spatial attention

Local features of a given sensor

<table>
<thead>
<tr>
<th>$x_{t,1}^i$</th>
<th>...</th>
<th>$x_{t,k}^i$</th>
<th>...</th>
<th>$x_{t,N_t}^i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{t-1}$</td>
<td>tanh</td>
<td>...</td>
<td>tanh</td>
<td>...</td>
</tr>
</tbody>
</table>

Importance of each local feature

<table>
<thead>
<tr>
<th>$a_1$</th>
<th>...</th>
<th>$a_k$</th>
<th>...</th>
<th>$a_{N_t}$</th>
</tr>
</thead>
</table>

Similarity matrix

- Global spatial attention

New global features at time $t$

<table>
<thead>
<tr>
<th>$y_t^1$</th>
<th>$y_t^k$</th>
<th>$y_{N_g}$</th>
</tr>
</thead>
</table>

Importance of each sensor

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>...</th>
<th>$\beta_k$</th>
<th>...</th>
<th>$\beta_{N_t}$</th>
</tr>
</thead>
</table>

Historical readings of each sensor
Temporal Attention

- Sequence-to-sequence learning architecture
- Select relevant previous time slots to make predictions
Evaluation

• Task 1 - water quality prediction
  – Water quality data
    • Residual chlorine
    • 10 kinds of time series
    • From 14 sensors in Shenzhen
    • Update each 5 minutes
  – Meteorology data
  – POIs data

• Task 2 - air quality prediction
  • Air quality data
    • PM2.5
    • 19 kinds of time series
    • From 35 sensors in Beijing
    • Hourly updates
  • Meteorology data
  • POIs data
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Water Quality</th>
<th></th>
<th>Air Quality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>ARIMA</td>
<td>8.61E-02</td>
<td>7.97E-02</td>
<td>31.07</td>
<td>20.58</td>
</tr>
<tr>
<td>VAR</td>
<td>5.02E-02</td>
<td>4.42E-02</td>
<td>24.60</td>
<td>16.17</td>
</tr>
<tr>
<td>GBRT</td>
<td>5.17E-02</td>
<td>3.30E-02</td>
<td>24.00</td>
<td>15.03</td>
</tr>
<tr>
<td>FFA</td>
<td>6.04E-02</td>
<td>4.10E-02</td>
<td>23.83</td>
<td>15.75</td>
</tr>
<tr>
<td>stMTMVL</td>
<td>6.07E-02</td>
<td>4.16E-02</td>
<td>29.72</td>
<td>19.26</td>
</tr>
<tr>
<td>stDNN</td>
<td>5.77E-02</td>
<td>3.99E-02</td>
<td>25.64</td>
<td>16.49</td>
</tr>
<tr>
<td>LSTM</td>
<td>6.89E-02</td>
<td>5.04E-02</td>
<td>24.62</td>
<td>16.70</td>
</tr>
<tr>
<td>Seq2seq</td>
<td>5.80E-02</td>
<td>4.03E-02</td>
<td>24.55</td>
<td>15.09</td>
</tr>
<tr>
<td>DA-RNN</td>
<td>5.02E-02</td>
<td>3.52E-02</td>
<td>24.25</td>
<td>15.17</td>
</tr>
<tr>
<td>GeoMAN</td>
<td><strong>4.34E-02</strong></td>
<td><strong>3.02E-02</strong></td>
<td><strong>22.86</strong></td>
<td><strong>14.08</strong></td>
</tr>
</tbody>
</table>

Visualization: Dynamic Correlation

- Case study over air quality dataset
  - Discuss on sensor $S_0$
  - 4:00 to 16:00 on Feb. 28, 2017

(a) Air quality stations in Beijing

(b) Plot of local spatial attention weights

(c) Plot of global spatial attention weights
基于大数据和AI的空气质量预测
Challenges

- Multiple influential factors with complex interactions
  - Pollution sources, direct factors and indirect factors
  - Affected by multiply factors simultaneously

- Dynamic spatio-temporal correlation and sudden changes
  - Urban air changes over location and time significantly
  - AQI drops very sharply in a very short time span
Deep Distributed Fusion Network

Spatial Transformation
- Air pollution dispersion
- Spatial correlation
- Scalability

Distributed FusionNet
- HW/WF/SP/MP nets to capture different individual influences
- Capture holistic influence (HI)

Weighted Merge
\[ \hat{y} = \text{Sigmoid}( y_{hw} \cdot w_{hw} + y_{wf} \cdot w_{wf} + y_{sp} \cdot w_{sp} + y_{mp} \cdot w_{mp} + y_{hi} \cdot w_{hi} ) \]

Xiuwen Yi, Junbo Zhang, et al. Deep Distributed Fusion Network for Air Quality Prediction. KDD
Official Prediction

- Advantages beyond Weather-Forecast-Based Method (WFM)
  - Spatial granularity: station vs district
  - Farther predictive capability: 48 vs 12 hours
  - Updating frequency: 1 hour vs 12 hours
  - Need less data sources
  - More accurate, 22% improvement

10/1/2014 to 12/30/2016.
Beijing Municipal Environmental Monitoring Center (using WFM)

<table>
<thead>
<tr>
<th>Method</th>
<th>Station Level</th>
<th>District Level</th>
<th>Update</th>
<th>Grained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>acc</td>
<td>mae</td>
<td>acc</td>
<td>mae</td>
</tr>
<tr>
<td>WFM</td>
<td>0.54</td>
<td>54.5</td>
<td>0.64</td>
<td>46.1</td>
</tr>
<tr>
<td>DeepAir</td>
<td>0.77</td>
<td>26.7</td>
<td>0.86</td>
<td>17.9</td>
</tr>
</tbody>
</table>
火力发电行业背景

- 发电形式
  - 水电，火电，核电，风电和太阳能
  - 火力发电约占总发电量70%

- 火电装机容量
  - 全国总10.5亿千瓦
  - 约2000台机组（锅炉）

假如发电效率从90%提升到90.5%

- 一台60万千瓦火电机组一年可节省煤
  - 0.55万吨 = 500万元

- 全国一年可节省煤
  - 962.5万吨 = 87.5亿元

经济效益

- 全北京四个月的总用煤量

环境效益

- 减少污染物排放
  - 二氧化硫：38.5万吨
  - 氮氧化物：48.1万吨

- 总治理成本：23.1亿元

每年为国家节约100个亿!
AI + 火力发电

离线学习

在线运行

用更少的煤

发更多的电

更少的污染
深度强化学习优化

状态变量：$s_i$

智能体

动作变量：$a_i$

传感器

控制

环境

策略：$s_t \rightarrow a_t^*$

<table>
<thead>
<tr>
<th>磨煤机出口压力</th>
<th>给煤机给煤量</th>
</tr>
</thead>
<tbody>
<tr>
<td>燃烧器风粉温度</td>
<td>冷/热风阀门</td>
</tr>
<tr>
<td>主蒸汽压力</td>
<td>混合风阀门</td>
</tr>
<tr>
<td>主蒸汽温度</td>
<td>减温水调节阀</td>
</tr>
<tr>
<td>炉膛负压</td>
<td>二次风C，F挡板开度</td>
</tr>
<tr>
<td>冷/热一次风量</td>
<td>燃尽风箱风门开度</td>
</tr>
<tr>
<td>给水流量</td>
<td>一次风机导叶位置</td>
</tr>
<tr>
<td>给水温度</td>
<td>送风机导叶位置</td>
</tr>
<tr>
<td>过量空气系数</td>
<td>引风机导叶位置</td>
</tr>
<tr>
<td>...</td>
<td>过热器烟气挡板</td>
</tr>
</tbody>
</table>
深度强化学习优化

**AlphaGo**

- Rollout policy
- SL policy network
- RL policy network
- Value network
- Policy network
- Value network

**Human expert positions**
- $p_\pi$
- $p_o$
- $p_p$

**Self-play positions**
- $r_\theta$

**Neural network**
- $p_{alp}(a|s)$
- $r_\theta(s')$
Take away message

- No universal approach
  - Highly dependent on data / problem / model
- Know your data
  - Develop highly customized models based on the property of the data
  - Important in model development!
- Incorporate domain knowledge if possible
  - Exploit structural properties in the data
  - Helpful especially when data is limited
- Get rid of assumptions!!!!
  - Assumptions are bad!
  - If any, all assumptions should come from data
Take away message

• Need a different toolbox

• Change your mindset!
  – Let the data speak for itself
  – Stay away from traditional way of thinking
What to expect

• New and hot area, lots of opportunities
  – Many open questions
  – Solve problems that are not solvable before!
  – Many things to be done, especially for traditional engineering fields

• Conduct high impact research
  – Work on really interesting problems

• Develop something really useful
  – Build practical, deployable real-time applications
  – Papers should not be the final outcome
We Are Hiring!