Deep Sequence Learning with Auxiliary Information for Traffic Prediction

Binbing Liao, Jingqing Zhang, Chao Wu, Douglas McIlwraith, Tong Chen, Shengwen Yang, Yike Guo and Fei Wu

Zhejiang University, Imperial College London and Baidu Inc.

Introduction

Traffic prediction is crucial to many applications including traffic network planning, route guidance, and congestion avoidance. However, there are several important factors that affect traffic prediction:

- Offline geographical and social factors. e.g., width, topologies, holidays.
- Online potential influence. e.g., online crowd map queries.
- Limited dataset. e.g., either small or not publicly available.

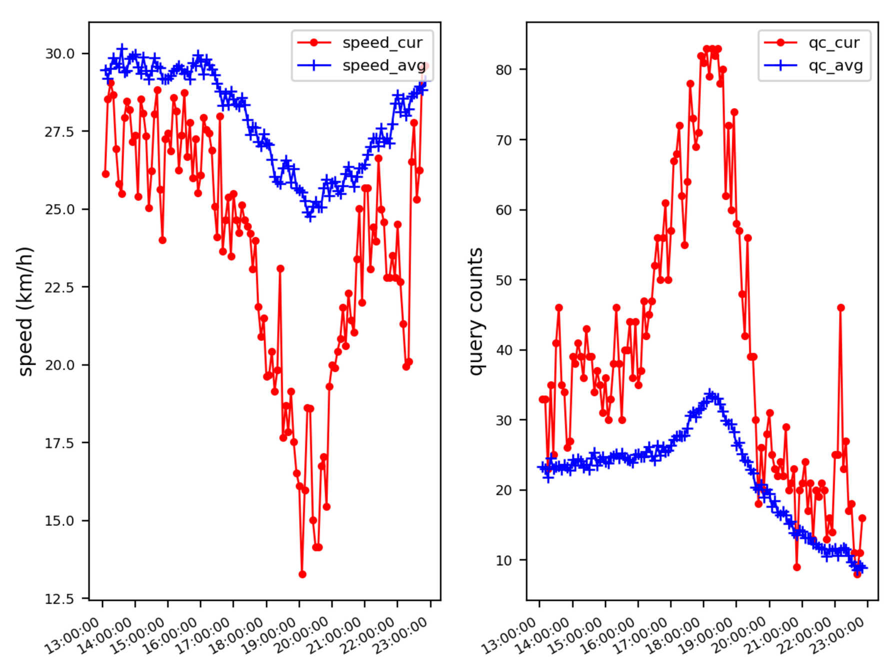

Figure 1: The traffic speed (left) and online query counts (right) around the Beijing Capital Gym on April 8, 2017. The red line denotes the unusual traffic speed (query counts) while the blue line indicates the usual traffic speed (query counts). At 19:00 PM, there was the Fish Leong Concert in the Capital Gym.

Q-traffic dataset

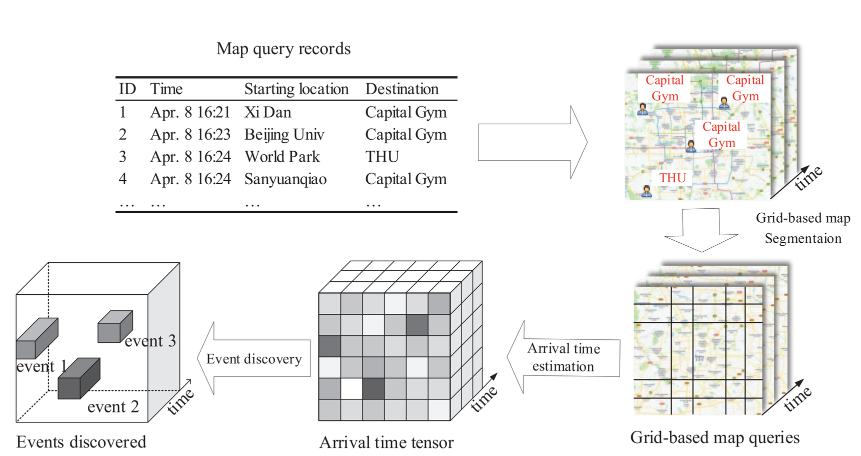

Figure 2: The flowchart of the mining of potential traffic in queries. A set of map queries is segmented into grids, then we estimate the arrival time at each query's destination, thus constructing an arrival time tensor. An event discovery algorithm is used to discover the events from the arrival time tensor.

Table 1: Comparison of different datasets for traffic prediction.

Datasets	Scale	Road info.	Road net.	Auxiliary info.	Highway	Urban	Available
Subset of PeMS							
State Route 22, Garden Grove	9						
PeMSD7 (S)	228		√		√		√
San Francisco Bay area	943						
PeMSD7 (L)	1,026						
Subset of Beijing							
Ring road around Beijing	2					√	
Beijing 4th ring road	3					√	
Beijing 2nd/3rd ring road	80		√		√		
Beijing 2nd/3rd ring road	122				√		
Bejing taxi dataset	236				√	√	
Bejing taxi dataset	352				√	√	
I-80 in California	6		√		√		√
Busan Metropolitan City	10		√			√	
California PATH	12				√		
Corridor in Orlando	71				√		
Rome dataset	120		√			√	
D100	122			weather		√	
Bedok area	226		√		√	√	
Los Angeles	1,642		√		√	√	
Los Angeles	4,048		√		√	√	
Dallas-Forth Worth area	4,764				√		
Subnetwork in Singapore	5,024		√		√	√	
Q-Traffic Dataset	15,073	√	√	map query	√	√	√

Methods

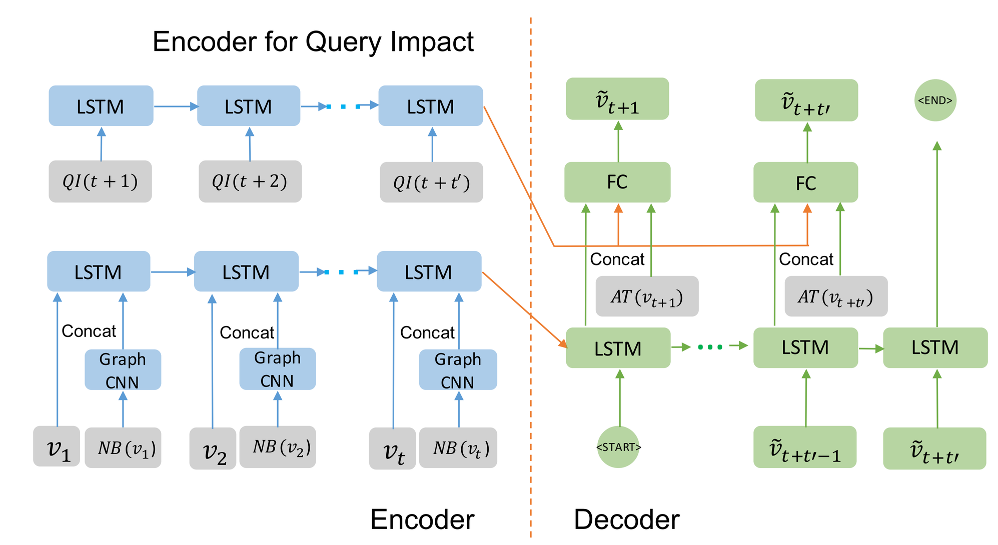

Figure 3: Hybrid model integrates all three auxiliary domains including attributes $\boldsymbol{AT}(v_t)$, spatial relation $\boldsymbol{NB}(v_t)$, and query impact $\boldsymbol{QI}(l,t)$.

Results

Table 2: Examples of discovered events.

Time	Grid	QC_cur	QC_last	Top1 query word	Top1_qc	Description
2017-04-08 14:00-20:00	(26, 39)	3431	417	Capital Gym	2724	Fish Leong Concert
2017-04-11 08:00-10:00	(24, 38)	447	93	Beijing Shangri-La Restaurant	304	IBM Data Scientist Forum
2017-04-15 08:00-16:00	(13, 47)	4551	2202	Beijing Botanical Garden	3849	Spring outing
2017-04-15 16:00-20:00	(21, 34)	2173	207	Letv sports center	1831	Chou Chuan-huing Concert
2017-04-30 08:00-18:00	(22, 47)	7283	3607	Summer Palace	7149	Summer Palace (May Day)
2017-04-30 08:00-18:00	(26, 46)	3691	1582	Tsinghua University	3102	106th Anniversary of THU

Table 3: Err_T (%): MAPE on the whole testing set.

Prediction	15-min	30-min	45-min	60-min	75-min	90-min	105-min	120-min	Overall
RF	6.00	9.15	10.20	10.66	10.98	11.21	11.39	11.56	10.14
SVR	5.44	9.20	10.07	10.34	10.51	10.65	10.76	10.83	9.73
Seq2Seq	4.61	8.22	9.28	9.72	9.98	10.27	10.48	10.61	9.23
Seq2Seq+AT	4.53	8.06	9.09	9.48	9.70	9.84	9.93	10.01	8.83
Seq2Seq+NB	**4.52**	8.05	9.07	9.45	9.67	9.83	9.93	9.99	8.81
Seq2Seq+QI	4.58	8.01	8.95	9.31	9.51	9.66	9.80	9.94	8.72
Hybrid	**4.52**	**7.93**	**8.89**	**9.24**	**9.43**	**9.56**	**9.69**	**9.78**	**8.63**

Table 4: Err_E (%): MAPE during events on the testing set.

Prediction	15-min	30-min	45-min	60-min	75-min	90-min	105-min	120-min	Overall
RF	6.14	9.51	10.81	11.45	11.84	12.13	12.38	12.56	10.85
SVR	5.64	9.56	10.59	11.02	11.32	11.56	11.73	11.83	10.41
Seq2Seq	4.76	8.52	9.87	10.52	10.91	11.31	11.60	11.80	9.91
Seq2Seq+AT	4.65	8.32	9.63	10.23	10.58	10.81	10.98	11.13	9.54
Seq2Seq+NB	4.63	8.25	9.53	10.10	10.45	10.70	10.89	11.02	9.45
Seq2Seq+QI	4.69	8.18	9.37	9.93	10.28	10.55	10.77	10.98	9.34
Hybrid	**4.61**	**8.09**	**9.30**	**9.84**	**10.16**	**10.39**	**10.60**	**10.76**	**9.22**

Contributions

- We release a large-scale traffic prediction dataset with offline and online auxiliary information including map crowd search queries, road intersections and geographical and social attributes.
- We integrate the sequence to sequence deep neural networks with geographical and social attributes via a wide and deep manner.
- To incorporate the spatial dependencies within local road network, we utilise the graph convolution neural network to embed the traffic speed of neighbouring road segments.
- We quantify the potential influence and devise a query impact algorithm to calculate the impact that online crowd queries have on the road segments.
- We propose a hybrid Seq2Seq model which incorporates the offline geographical and social attributes, spatial dependencies and online crowd queries with a deep fusion.

Acknowledgements

We would like to thank the support from Baidu Map and ZJU-Imperial Joint Lab. This work was supported by the National Basic Research Program (973) of China (No. 2015CB352302) and the National Natural Science Foundation of China (Nos. 61625107, U1611461, U1509206). Jingqing Zhang was funded by LexisNexis HPCC Systems Academic Program.

Contact Information

- Code & dataset: https://github.com/JingqingZ/BaiduTraffic
- Email: bbliao@zju.edu.cn & jingqing.zhang15@imperial.ac.uk
- Phone: +86 181 6715 3400

Imperial College London

Baidu 百度