LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai,
Chao Huang, Zhenguang Liu, Bryan Hooi, Roger Zimmermann

Background

- Road traffic forecasting is one of the most critical components of Intelligent Transportation Systems
  - Urban planning
  - Traffic management
  - Public safety

- Spatio-temporal graph neural networks are the widely embraced tools for accurate traffic predictions
  - Graph neural networks to capture spatial correlations
  - Sequential models to model temporal dependencies

Motivation

- Historically, high-quality and large-scale benchmark datasets have proven their value in driving research frontiers
  - ImageNet in computer vision
  - GLUE in natural language processing
  - OGB in the general graph

- In traffic forecasting, we argue that *commonly-used datasets present critical issues that may pose obstacles to future progress*

Issues of Existing Datasets

Released Source	Dataset	Nodes	Edges	Degree	Meta	Time Range	Frames	Data Points
Yu et al. [34]	PeMSD7(M)	228	1,664	7.3	6	05/01/2012 – 06/30/2012	12,672	2.89M
	PeMSD7(L)	1,026	14,534	14.2	0	05/01/2012 – 06/30/2012	12,672	13.00M
Li et al. [19]	METR-LA	207	1,515	7.3	3	03/01/2012 – 06/27/2012	34,272	7.09M
	PEMS-BAY	325	2,369	7.3	3	01/01/2017 – 06/30/2017	52,116	16.94M
Song et al. [30]	PEMS03	358	546	1.5	1	09/01/2018 – 11/30/2018	26,208	9.38M
	PEMS04	307	338	1.1	0	01/01/2018 – 02/28/2018	16,992	5.22M
	PEMS07	883	865	1.0	0	05/01/2017 – 08/06/2017	28,224	24.92M
	PEMS08	170	276	1.6	0	07/01/2016 – 08/31/2016	17,856	3.04M
LargeST (ours)	CA	8,600	201,363	23.4	9	01/01/2017 – 12/31/2021	525,888	4.52B
	GLA	3,834	98,703	25.7	9	01/01/2017 – 12/31/2021	525,888	2.02B
	GBA	2,352	61,246	26.0	9	01/01/2017 – 12/31/2021	525,888	1.24B
	SD	716	17,319	24.2	9	01/01/2017 – 12/31/2021	525,888	0.38B

- Existing datasets: comprising merely hundreds of nodes and edges
- LargeST: including up to 50.6x more nodes and 729.6x more edges

Issues of Existing Datasets

Released Source	Dataset	Nodes	Edges	Degree	Meta	Time Range	Frames	Data Points
Yu et al. [34]	PeMSD7(M)	228	1,664	7.3	6	05/01/2012 – 06/30/2012	12,672	2.89M
	PeMSD7(L)	1,026	14,534	14.2	0	05/01/2012 – 06/30/2012	12,672	13.00M
Li et al. [19]	METR-LA	207	1,515	7.3	3	03/01/2012 – 06/27/2012	34,272	7.09M
	PEMS-BAY	325	2,369	7.3	3	01/01/2017 – 06/30/2017	52,116	16.94M
Song et al. [30]	PEMS03	358	546	1.5	1	09/01/2018 – 11/30/2018	26,208	9.38M
	PEMS04	307	338	1.1	0	01/01/2018 – 02/28/2018	16,992	5.22M
	PEMS07	883	865	1.0	0	05/01/2017 – 08/06/2017	28,224	24.92M
	PEMS08	170	276	1.6	0	07/01/2016 – 08/31/2016	17,856	3.04M
LargeST (ours)	CA	8,600	201,363	23.4	9	01/01/2017 – 12/31/2021	525,888	4.52B
	GLA	3,834	98,703	25.7	9	01/01/2017 – 12/31/2021	525,888	2.02B
	GBA	2,352	61,246	26.0	9	01/01/2017 – 12/31/2021	525,888	1.24B
	SD	716	17,319	24.2	9	01/01/2017 – 12/31/2021	525,888	0.38B

- Existing datasets: ofter spanning less than 6 months of data
- LargeST: covering an unprecedented 5 years of data

Issues of Existing Datasets

Released Source	Dataset	Nodes	Edges	Degree	Meta	Time Range	Frames	Data Points
Yu et al. [34]	PeMSD7(M)	228	1,664	7.3	6	05/01/2012 – 06/30/2012	12,672	2.89M
	PeMSD7(L)	1,026	14,534	14.2	0	05/01/2012 – 06/30/2012	12,672	13.00M
Li et al. [19]	METR-LA	207	1,515	7.3	3	03/01/2012 – 06/27/2012	34,272	7.09M
	PEMS-BAY	325	2,369	7.3	3	01/01/2017 – 06/30/2017	52,116	16.94M
Song et al. [30]	PEMS03	358	546	1.5	1	09/01/2018 – 11/30/2018	26,208	9.38M
	PEMS04	307	338	1.1	0	01/01/2018 – 02/28/2018	16,992	5.22M
	PEMS07	883	865	1.0	0	05/01/2017 – 08/06/2017	28,224	24.92M
	PEMS08	170	276	1.6	0	07/01/2016 – 08/31/2016	17,856	3.04M
LargeST (ours)	CA	8,600	201,363	23.4	9	01/01/2017 – 12/31/2021	525,888	4.52B
	GLA	3,834	98,703	25.7	9	01/01/2017 – 12/31/2021	525,888	2.02B
	GBA	2,352	61,246	26.0	9	01/01/2017 – 12/31/2021	525,888	1.24B
	SD	716	17,319	24.2	9	01/01/2017 – 12/31/2021	525,888	0.38B

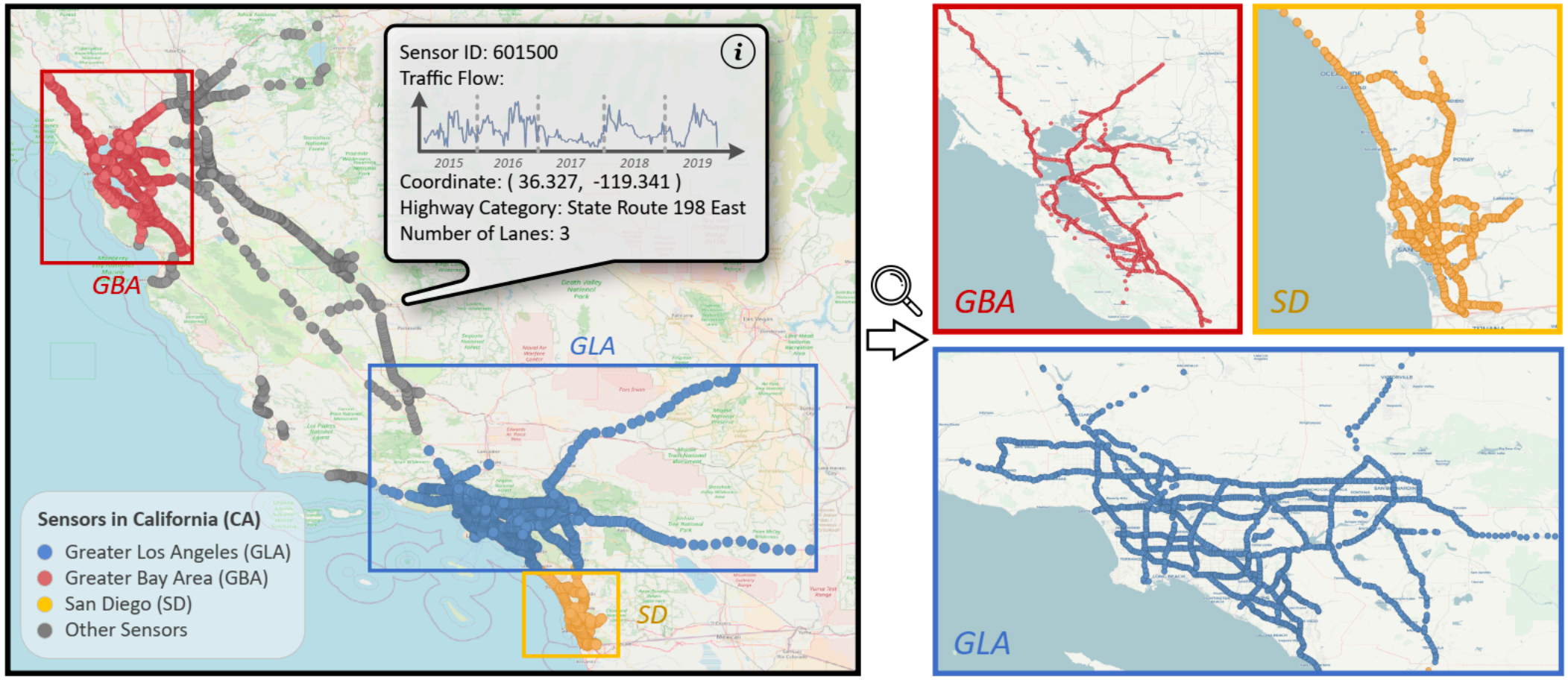
- Existing datasets: insufficient metadata available for individual nodes
- LargeST: comprising comprehensive node metadata

The LargeST Benchmark

- To this end, we propose **LargeST** as a new benchmark dataset, with the goal of *facilitating the development of accurate and efficient methods in the context of large-scale traffic forecasting*

- In this work, we also
  - Conduct comprehensive data analysis
  - Implement a suite of well-known baselines
  - Perform extensive benchmarking experiments

Visual Overview

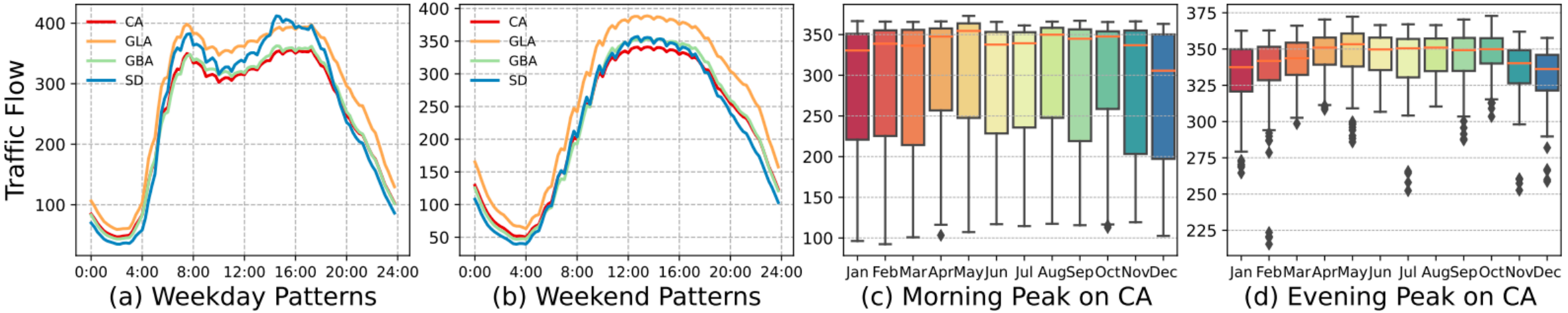


(a) Overview of the LargeST dataset

(b) Fine-grained distribution of sensors

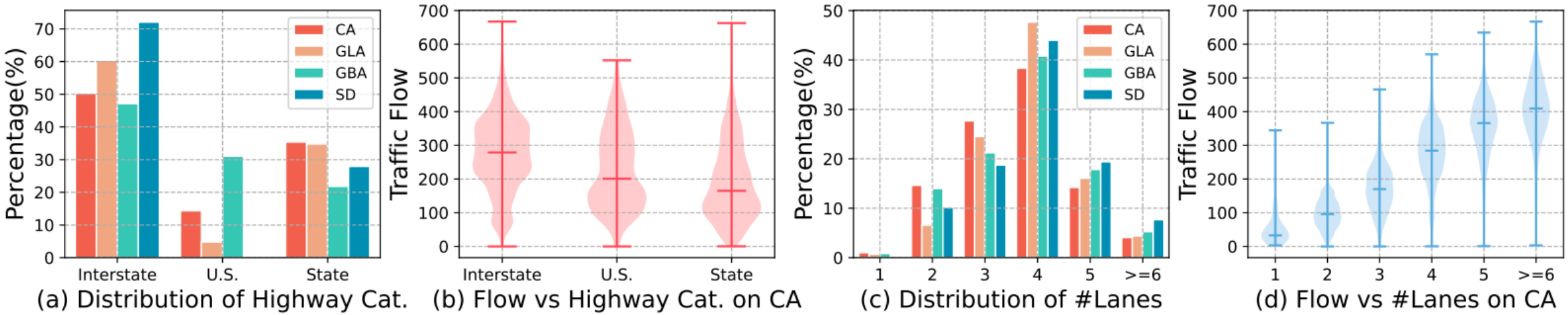
Data Analysis

- Relations between traffic flow and the factors in space (regional disparities) and time (temporal dynamics)



Data Analysis

- Relations between traffic flow and two crucial meta features: the highway categories and the number of lanes



Experiments

- Recent methods like DGCRN and D2STGNN demonstrate impressive performance on the datasets of SD and GBA
- Competitive methods introduced 3-4 years ago, namely GWNET and AGCRN, continue to perform well

Data	Method	Param	Horizon 3			Horizon 6			Horizon 12			Average		
			MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
SD	HL	–	33.61	50.97	20.77%	57.80	84.92	37.73%	101.74	140.14	76.84%	60.79	87.40	41.88%
	LSTM	98K	19.03	30.53	11.81%	25.84	40.87	16.44%	37.63	59.07	25.45%	26.44	41.73	17.20%
	DCRNN	373K	17.14	27.47	11.12%	20.99	33.29	13.95%	26.99	42.86	18.67%	21.03	33.37	14.13%
	AGCRN	761K	15.71	27.85	11.48%	18.06	31.51	13.06%	21.86	39.44	16.52%	18.09	32.01	13.28%
	STGCN	508K	17.45	29.99	12.42%	19.55	33.69	13.68%	23.21	41.23	16.32%	19.67	34.14	13.86%
	GWNET	311K	15.24	25.13	9.86%	17.74	29.51	11.70%	**21.56**	**36.82**	15.13%	**17.74**	29.62	11.88%
	ASTGCN	2.2M	19.56	31.33	12.18%	24.13	37.95	15.38%	30.96	49.17	21.98%	23.70	37.63	15.65%
	STTN	114K	16.22	26.22	10.63%	18.76	30.98	12.80%	22.62	39.09	16.14%	18.69	31.11	12.82%
	STGODE	729K	16.75	28.04	11.00%	19.71	33.56	13.16%	23.67	42.12	16.58%	19.55	33.57	13.22%
	DSTAGNN	3.9M	18.13	28.96	11.38%	21.71	34.44	13.93%	27.51	43.95	19.34%	21.82	34.68	14.40%
	DGCRN	243K	15.34	25.35	10.01%	18.05	30.06	11.90%	22.06	37.51	15.27%	18.02	30.09	12.07%
	D^2STGNN	406K	**14.92**	**24.95**	**9.56%**	**17.52**	**29.24**	**11.36%**	22.62	37.14	**14.86%**	17.85	**29.51**	**11.54%**
GBA	HL	–	32.57	48.42	22.78%	53.79	77.08	43.01%	92.64	126.22	92.85%	56.44	79.82	48.87%
	LSTM	98K	20.38	33.34	15.47%	27.56	43.57	23.52%	39.03	60.59	37.48%	27.96	44.21	24.48%
	DCRNN	373K	18.71	30.36	14.72%	23.06	36.16	20.45%	29.85	46.06	29.93%	23.13	36.35	20.84%
	AGCRN	777K	18.31	30.24	14.27%	21.27	34.72	16.89%	**24.85**	**40.18**	20.80%	21.01	34.25	16.90%
	STGCN	1.3M	21.05	34.51	16.42%	23.63	38.92	18.35%	26.87	44.45	21.92%	23.42	38.57	18.46%
	GWNET	344K	17.85	29.12	13.92%	21.11	**33.69**	17.79%	25.58	40.19	23.48%	20.91	**33.41**	17.66%
	ASTGCN	22.3M	21.46	33.86	17.24%	26.96	41.38	24.22%	34.29	52.44	32.53%	26.47	40.99	23.65%
	STTN	218K	18.25	29.64	14.05%	21.06	33.87	17.03%	25.29	40.58	21.20%	20.97	33.78	16.84%
	STGODE	788K	18.84	30.51	15.43%	22.04	35.61	18.42%	26.22	42.90	22.83%	21.79	35.37	18.26%
	DSTAGNN	26.9M	19.73	31.39	15.42%	24.21	37.70	20.99%	30.12	46.40	28.16%	23.82	37.29	20.16%
	DGCRN	374K	18.02	29.49	14.13%	21.08	34.03	16.94%	25.25	40.63	21.15%	20.91	33.83	16.88%
	D^2STGNN	446K	**17.54**	**28.94**	**12.12%**	**20.92**	33.92	**14.89%**	25.48	40.99	**19.83%**	**20.71**	33.65	**15.04%**

Experiments

- The complex model designs of DGCRN and D2STGNN prevent them from scaling to larger datasets: GLA and CA

GLA	HL	–	33.66	50.91	19.16%	56.88	83.54	34.85%	98.45	137.52	71.14%	59.58	86.19	38.76%
	LSTM	98K	20.02	32.41	11.36%	27.73	44.05	16.49%	39.55	61.65	25.68%	28.05	44.38	17.23%
	DCRNN	373K	18.41	29.23	10.94%	23.16	36.15	14.14%	30.26	46.85	19.68%	23.17	36.19	14.40%
	AGCRN	792K	**17.27**	29.70	10.78%	**20.38**	34.82	**12.70%**	**24.59**	42.59	**16.03%**	**20.25**	34.84	**12.87%**
	STGCN	2.1M	19.86	34.10	12.40%	22.75	38.91	14.11%	26.70	45.78	17.00%	22.64	38.81	14.17%
	GWNET	374K	17.28	**27.68**	**10.18%**	21.31	**33.70**	13.02%	26.99	**42.51**	17.64%	21.20	**33.58**	13.18%
	ASTGCN	59.1M	21.89	34.17	13.29%	29.54	45.01	19.36%	39.02	58.81	29.23%	28.99	44.33	19.62%
	STGODE	841K	18.10	30.02	11.18%	21.71	36.46	13.64%	26.45	45.09	17.60%	21.49	36.14	13.72%
	DSTAGNN	66.3M	19.49	31.08	11.50%	24.27	38.43	15.24%	30.92	48.52	20.45%	24.13	38.15	15.07%
CA	HL	–	30.72	46.96	20.43%	51.56	76.48	37.22%	89.31	125.71	76.80%	54.10	78.97	41.61%
	LSTM	98K	19.04	31.28	13.19%	26.49	42.63	19.57%	38.22	60.29	30.28%	26.89	43.11	20.16%
	DCRNN	373K	17.55	28.21	12.68%	21.79	34.27	16.67%	28.56	44.34	23.84%	21.87	34.41	17.06%
	STGCN	4.5M	18.99	32.37	14.84%	21.37	36.46	**16.27%**	**24.94**	**42.59**	**19.74%**	21.33	36.39	**16.53%**
	GWNET	469K	**17.14**	**27.81**	**12.62%**	21.68	**34.16**	17.14%	28.58	44.13	24.24%	21.72	**34.20**	17.40%
	STGODE	1.0M	17.57	29.91	13.91%	**20.98**	36.62	16.88%	25.46	45.99	21.00%	**20.77**	36.60	16.80%

Experiments

Method	SD				GBA				GLA				CA			
	BS	Train	Infer	Total	BS	Train	Infer	Total	BS	Train	Infer	Total	BS	Train	Infer	Total
LSTM	64	21	6	1	64	115	17	4	64	188	29	6	32	415	61	13
DCRNN	64	867	150	28	64	1,816	319	59	43	2,491	435	81	19	4,845	851	158
AGCRN	64	92	15	3	64	536	83	17	45	1,413	245	46	–	–	–	–
STGCN	64	53	16	2	64	160	54	6	64	268	86	10	64	701	206	25
GWNET	64	97	14	3	64	483	66	15	64	1,028	139	32	44	4,105	548	113
ASTGCN	64	128	19	4	45	1,126	147	35	17	3,060	393	77	–	–	–	–
STTN	64	208	26	6	7	1,758	197	50	–	–	–	–	–	–	–	–
STGODE	64	188	26	6	49	710	103	23	30	1,305	192	42	13	4,212	659	135
DSTAGNN	64	240	23	7	27	1,959	171	53	10	5,241	467	120	–	–	–	–
DGCRN	64	430	76	14	12	4,461	605	138	–	–	–	–	–	–	–	–
D^2STGNN	45	563	69	14	4	5,885	796	148	–	–	–	–	–	–	–	–

- The CA dataset poses significant challenges for existing traffic forecasting models, as only half of the selected baselines are capable of running on it.

Future Opportunities

- According to our thorough data analysis and extensive experiment results, we highlight the following opportunities in future research

➡The utilization of spatial, temporal, and metadata features

➡A valuable testbed for the challenges of temporal distribution shifts

➡The development of simple yet effective methods

➡The development of foundation forecasting models

THANK YOU

Official Code Repository
https://github.com/liuxu77/LargeST