Robust Spatio-Temporal Centralized Interaction for OOD Learning

Jiaming Ma[1] Bingwu Wang*,[1,2] Pengkun Wang[1,2] Zhengyang Zhou[1,2] Xu Wang[1,2] Yang Wang *,[1,2]

[1]University of Science and Technology of China (USTC), Hefei, China [2]Suzhou Institute for Advanced Research, USTC, Suzhou, China

Abstract

Recently, spatio-temporal graph convolutional networks have achieved dominant performance in spatio-temporal prediction tasks. However, most models relying on node-to-node messaging interaction exhibit sensitivity to spatio-temporal shifts, encountering out-of-distribution (OOD) challenges. To address these issues, we introduce **S**patio-**T**emporal **O**OD **P**rocessor (STOP), which employs a centralized messaging mechanism along with a message perturbation mechanism to facilitate robust spatio-temporal interactions. Specifically, the centralized messaging mechanism integrates Context-Aware Units for coarse-grained spatio-temporal feature interactions with nodes, effectively blocking traditional node-to-node messages. We also implement a message perturbation mechanism to disrupt this messaging process, compelling the model to extract generalizable contextual features from generated variant environments. Finally, we customize a spatio-temporal distributionally robust optimization approach that exposes the model to challenging environments, thereby further enhancing its generalization capabilities. Compared with 14 baselines across six datasets, STOP achieves up to 17.01% improvement in generalization performance and 18.44% improvement in inductive learning performance.

Introduction

The encouraging success of STGNNs is predicated on the IID assumption. In reality, the distributional characteristics or graph structures of spatio-temporal data evolve over time, presenting OOD generalization challenges for STGNNs.

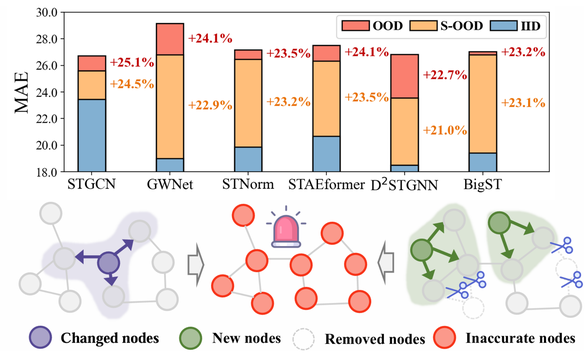

Figure 1. Spatiotemporal OOD.

With the LargeST-SD dataset as an example, we report the performance of advanced STGNNs in both IID and OOD scenarios, as shown in Figure 1. (a). The results indicate that their performance rapidly deteriorates when facing spatio-temporal OOD challenges, particularly in structural shift (S-OOD) scenario. Embarrassingly, the models' reliance on node-to-node messaging appears to hinder their effectiveness. Ablation experiments further support this: for some advanced STGNNs, variants without node-to-node messaging mechanism perform embarrassingly better. This is because the knowledge learned acquired through this mechanism is coupled with the features of the training graph, and this knowledge is difficult to generalize to unseen graphs during testing. As illustrated in Figure 1. (b), when node features change, STGNNs struggle to accurately represent these nodes. Furthermore, when certain nodes disappear from the graph, their neighbors are significantly impacted, as they can no longer aggregate information along the trained paths. This propagation of errors through the node-to-node messaging mechanism adversely affects the accuracy of the entire graph representation. On the other hand, generating accurate representations for new nodes, i.e., inductive learning, also poses a significant challenge for STGNNs.

Methdology
A Spatiotemporal OOD Processor

Our model consists of three parts: it incorporates a robust centralized messaging mechanism and a message perturbation mechanism.

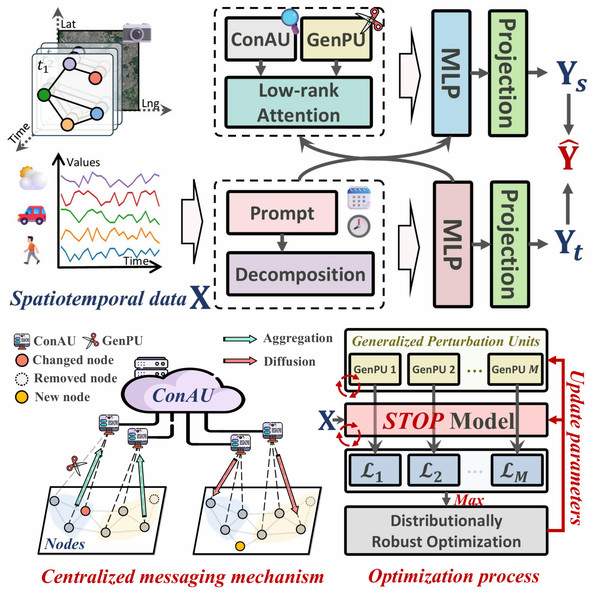

Figure 2. Overall architecture (upper) and centralized messaging mechanism for robust spatio-temporal interaction.

Centralized messaging

It constrains nodes to interact exclusively with Context Aware Units (ConAU) for feature interaction, thereby enhancing the resilience of model to spatio-temporal shifts. For the i-th head, the calculation of low-rank attention is as follows,

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \underbrace{\text{Softmax}(\alpha\mathbf{Q}\mathbf{K}^\top)}_{\textit{Diffusion}} \times \underbrace{\text{Softmax}(\alpha\mathbf{K}\mathbf{Q}^\top)\mathbf{V}}_{\textit{Aggregation}}.$$

Message perturbation

It is equipped with Generalized Perturbation Units (GenPU), disrupts node interactions with ConAU and includes a specialized spatio-temporal distributionally robust optimization (DRO) for GenPU, facilitating the model's acquisition of causal knowledge across diverse environments.

$$\tilde{\mathcal{A}}(\mathbf{Q},\mathbf{K},\mathbf{V};\mathbf{G}_i) = \text{Softmax}(\alpha\mathbf{Q}\mathbf{K}^\top) \times \underbrace{\text{Softmax}(\alpha\mathbf{K}\mathbf{Q}^\top + \mathbf{G}_i)\mathbf{V}}_{\textit{Perturbation operation}}.$$

Spatiotemporal DRO

The final optimization objective is defined as follows,

$$\min_{f} \sup_{\boldsymbol{g}\in\mathbb{R}^N} \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim(\mathcal{X},\mathcal{Y})}[\mathcal{L}(f(\mathrm{X}),\mathrm{Y};\boldsymbol{g})], \qquad s.t.\ \|\boldsymbol{g}\|_0 = s \in (0,N).$$

In fact, our proposed objective belongs to the DRO paradigm, which theoretically has superior generalization compared to the Empirical Risk Minimization (ERM) paradigm followed by most spatio-temporal models. ERM optimizes the model using only single training environment.

Experiment

Dataset settings: For the evaluation of temporal shift, we train the models using data from the first year and test them on each subsequent year. The training set comprises the first 60% of data from the initial year dataset, while the following 20% of data is used as the validation set. In each subsequent year, the last 20% of data is designated as the test set. This setup aims to accentuate the temporal distribution difference between the test and training sets, while maintaining a ratio of approximately 6:2:2 for the training, validation, and test sets. Regarding structural shift evaluation, we select a subset of nodes for training and validation. In the test set, we decrease the number of nodes by 10% and introduce 30% new nodes to simulate shifts in the graph structure and scale.

Table 1. OOD performance comparisons on selective datasets. The unit of MAPE is percent(%).We bold the best-performing model results in **red** and underline the sub-optimal model results in blue.

		Method	**Ours**	STONE	CaST	RPMixer	BigST	D²STGNN	STNN	STAEformer	STID	STNorm	GWNet	STGCN
SD	3	MAE	**17.71**	18.44	21.35	24.92	18.56	18.70	36.46	18.70	19.68	18.82	20.15	18.68
		RMSE	**28.45**	29.55	33.28	39.88	29.93	29.31	56.84	28.97	29.56	30.06	31.34	29.61
		MAPE	**11.73**	12.32	16.04	15.63	12.18	13.04	26.91	12.62	13.18	12.82	14.44	12.92
	6	MAE	**23.62**	25.10	29.28	42.37	25.66	25.13	36.91	25.80	25.87	26.00	28.07	25.25
		RMSE	**37.71**	39.66	45.24	66.45	40.61	38.77	57.59	40.73	40.86	41.20	43.00	39.48
		MAPE	**15.99**	17.56	21.49	26.15	18.03	17.46	27.15	17.59	18.03	18.03	21.17	17.34
	12	MAE	**32.59**	37.12	42.40	77.31	37.89	36.35	41.69	37.17	38.30	38.08	39.75	36.15
		RMSE	**51.82**	54.60	64.05	115.62	58.74	53.60	64.99	57.81	59.40	59.24	61.08	55.74
		MAPE	**22.89**	25.90	31.73	49.48	27.12	25.98	31.32	27.07	26.90	27.89	31.46	26.41
GBA	3	MAE	**18.33**	20.19	21.85	24.79	19.92	19.10	40.61	20.91	19.09	20.86	20.65	21.49
		RMSE	**29.70**	33.65	34.32	39.59	32.33	32.64	60.07	33.59	31.40	32.92	32.21	33.57
		MAPE	**13.64**	15.10	18.61	17.06	14.75	14.29	33.77	14.93	14.36	16.00	15.70	14.79
	6	MAE	**24.75**	25.84	29.70	40.77	28.64	26.10	40.50	28.61	26.90	31.24	28.39	30.05
		RMSE	**38.48**	41.96	45.16	62.24	43.93	41.72	59.96	44.03	42.15	46.69	42.60	44.97
		MAPE	**20.48**	21.24	25.77	29.48	22.25	21.26	33.68	22.41	21.79	25.57	22.74	22.84
	12	MAE	**34.93**	39.56	42.60	72.51	42.87	36.26	44.62	41.68	39.36	45.73	39.61	43.29
		RMSE	**53.10**	56.18	63.33	104.93	63.06	56.23	65.61	62.28	59.60	65.62	58.33	62.34
		MAPE	**31.09**	32.18	36.88	56.28	34.52	32.23	38.28	34.99	33.43	41.02	33.67	35.23

Hyperparameter Sensitivity Analysis: We analyze the sensitivity of the numer of ConAU and GenPU on the SD (upper) and KnowAir (lower) datasets on the Figure \ref{hyper}. When the number of ConAU K is set to 8 in SD dataset and 4 in KnowAir dataset. When K exceeds this value, the model creates too many ConAU, making it unable to focus on extracting invariant contextual features, thus introducing noise. When K is less than this value, too few perception units fail to learn sufficient invariant knowledge, leading to a decrease in the model's generalization performance. The number of GenPU M is set to 3 in SD dataset and 4 in KnowAir dataset. A smaller M may not provide sufficient training environment diversity, resulting in performance degradation. On the other hand, an excessive number of GenPU does not necessarily improve performance. Too large M means that the generated environment is too complex, which increases the learning difficulty of the model to extract causal knowledge.

Ablation Study: As illustrated in Figure~\ref{ablation}, the results show that each component of STOP helps to improve the OOD generalization. "w/o $\mathbf{Y}$" achieves poor prediction performance, which proves that the proposed collaborative component is effective for OOD. "w/o ConAU" removes ConAU and achieves high errors, proving that spatio-temporal interaction is also necessary in OOD scenarios. "w/o GenPU" has higher prediction errors because GenPU can help the model extract causal knowledge and enhance model robustness.

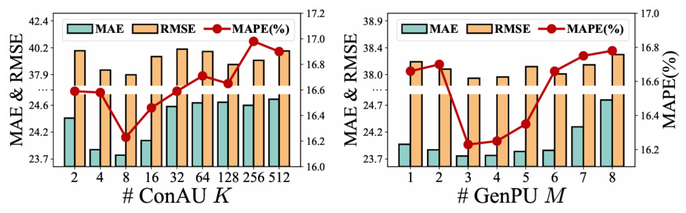

Figure 3. Hyper parameter experiments for K and M on SD dataset.

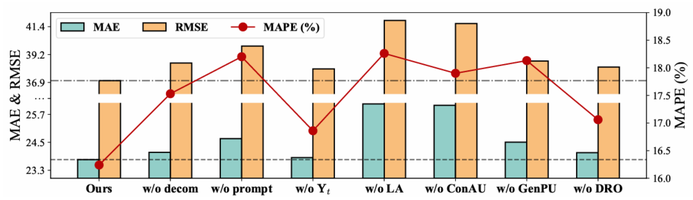

Figure 4. Ablation experiments on SD dataset.

Connection

Acknowledgements: This paper is partially supported by the National Natural Science Foundation of China (No.12227901). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

Personal Homepage: https://poorotterbob.github.io.

Available Code: https://github.com/PoorOtterBob/STOP.

Contact Emails: JiamingMa@mail.ustc.edu.cn.

ICML International Conference On Machine Learning