12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Causal-Transformer: Spatial-temporal causal attention-based transformer for time series prediction

Yaqi Zhu * **Fan Yang**† * **Andrei Torgashov** **

* *Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*
** *Process Control Laboratory, Institute of Automation and Control Processes, Far-Eastern Branch, Russian Academy of Sciences, Vladivostok, 690041 Russia*
(Corresponding author†, e-mail: zhuyq21@mails.tsinghua.edu.cn, yangfan@tsinghua.edu.cn, torgashov@iacp.dvo.ru)

Abstract: Real-time monitoring and accurate prediction of key variables are indispensable to ensure industrial production activities proceed as expected. With the increase in measurement data volume and the improvement of hardware computing power, the Transformer and its variants, due to their excellent capability in extracting global dependencies, are playing an increasingly important role among deep learning-based multidimensional time series prediction models. In addition, from the perspective of causality, cause variables contain parts of information in effect variables and can reduce the uncertainty of effect variables, which is beneficial for prediction. However, there has been relatively limited research on combining the Transformer and causal feature analysis. To fully use both advantages, this paper introduces the Causal-Transformer (CT) model, which utilizes semi-orthogonal projection to extract causal features from multiple input variables. A multi-head spatial-temporal causal attention mechanism is designed in the encoder block based on the classical Transformer model to simultaneously reduce feature dimensions and extract implicit causal features in both the temporal and spatial dimensions. The CT also utilizes the Granger causality analysis to select the causal teaching indicators of target variables to provide stable assistance by injecting explicit causality into the inputs of the decoder block. By leveraging more condensed and independent causal features, the CT possesses inherent advantages in predicting time series variables. Case study results show that the CT model outperforms the other models on the diesel refinery dataset, especially with a reduction of 46.0% and 30.4% in MSE towards the classic Transformer and informer in five-step prediction.

Keywords: Machine learning, Time series modelling, Transformer, Causal analysis, Attention machenism

1. INTRODUCTION

With the development trend of complexity, digitization, and high precision in industrial production processes, real-time monitoring and assessment of industrial processes play a crucial role in ensuring production safety and enhancing production efficiency (Kashpruk et al., 2023). In the era of the Industry 4.0 revolution, the rapid development of the Internet of Things (Mahdavinejad et al., 2018) and Cloud Computing (Chen et al., 2019) makes it possible to utilize multifarious sensors, leading to an increasing amount of multidimensional time series data generated during monitoring. So, accurate prediction of target time series variables reflecting product quality or physicochemical properties is undoubtedly beneficial for predictive maintenance, production optimization, anomaly detection, and other related tasks.

Multidimensional time series prediction methods can be mainly categorized into statistical regression models (Liu et al., 2016; Li et al., 2022), traditional machine learning-based models (Han et al., 2016; Sapankevych and Sankar, 2009), and deep learning-based models (Lim and Zohren, 2021). With the advancement of big data technology and hardware computing power, deep learning techniques have demonstrated significant potential in modeling complex time series data. Various sophisticated time-series prediction models, particularly the Transformer-like model (Geng et al., 2022), have been proposed in recent years.

Since the introduction of the Transformer, a series of novel Transformer-based models have been designed to address problems such as local feature extraction and memory storage, incorporating various attention mechanisms, local feature extraction modules, prior temporal feature

injection modules, and innovative model architectures. Li et al. (2019) adopted LogSparse self-attention and causal convolution to improve the local feature extraction and reduce memory cost. Also, Zhou et al. (2021) proposed that the informer mainly involved the ProbSparse self-attention and distilling operation to reduce time and space complexity. Shen and Wang (2022) borrowed ideas from the field of computer vision to introduce the CSPAttention and dilated causal convolution for exponentially receptive field growth. Attention mechanisms mentioned above chose the length and the dimension of queries separately rather than simultaneously selecting both. However, Zhang and Yan (2022) proposed the Two-Stage attention and applied multi-head attention mechanism to both dimensions, which implies that feature extraction of variable and temporal dimensions in Transformer-based time series prediction models is also a crucial problem to solve.

To extract the features from variable and temporal dimensions simultaneously in the industrial process, Yuan et al. (2021) proposed a spatiotemporal attention mechanism based on the LSTM encoder-decoder, enabling the model to focus on what is more relevant to target variables at different time steps. However, few works in time series prediction have combined spatial and temporal feature extraction with Transformer-like models.

Generally, a root cause is the fundamental reason leading to the deviation of a system from its normal state or the occurrence of a failure. So root cause analysis involves identifying how an inevitable failure occurred and discerning its reasons, which is significant in subsequent fault handling. Pearl et al. (2000) categorized causal relationships into three distinct levels: predictions, interventions, and counterfactuals. As one of the data-driven methods of causal analysis, Granger causality analysis elucidates the causality between variables from the prediction perspective by employing regression equations with lagged variables. It is believed that cause variables contain a portion of information from the effect variable, so cause variables can enhance predictive performance (Yu et al., 2022), reduce uncertainty, and influence the conditional probability of the effect variable.

Due to the benefits of causality for prediction tasks, it is natural to utilize Granger causality to select input variables of prediction models (Dong et al., 2019). Considering the remarkable modeling capability of the Transformer for non-linear dynamic relationships and the inherent superiority of causality in predictive tasks, integrating causality with the Transformer-like models can enhance the predictive performance (Liu et al., 2023). However, most of these works introduce causality by utilizing the causal convolution layers, which can be regarded as a relatively loose integration that prevents the leakage of future information. Causality can be integrated more tightly with the prediction models. Therefore, this paper strives to integrate causality and Transformer-based predictive models deeply by designing an attention mechanism that can capture causal features and reflect implicit causality and introducing the Granger causality into the architecture of Transformer-based models to seek direct guidance on causality. In light of these ideas, this paper proposes a novel model called Causal-Transformer (CT).

The remainder of the paper is organized as follows: Section 2 introduces the CT and the detailed architecture. In Section 3, the superiority of the CT is proved by comparing different models and conducting ablation experiments. Finally, a conclusion is provided in Section 4.

2. METHODOLOGY

2.1 Causal teaching indicators

The causal analysis can reveal the interdependent relationship between variables, which is determined by the system's physical structure and information flow and is impervious to external environmental disturbances. Therefore, using cause variables to predict effect variables, also called target variables, has a natural advantage. In terms of evaluating the causal relationship between time-series variables, Granger causality is a causal relationship that determines whether introducing a potential cause variable can improve the predictive result of the effect variable by comparing the prediction error of restricted and unrestricted regression equations.

The decoder of the classic Transformer takes the target variables as input and uses the teacher-forcing mechanism during training to predict target variables. Although this approach speeds up convergence, it does not utilize all available data. Hence, as the input to the encoder, partial easy-to-measure process variables having direct or indirect causal relationships with target variables are selected as part of the decoder's inputs as causal teaching indicators (CTIs). The CTIs can describe the inducing factors that lead to upcoming changes in the target variables, while the historical target variables provide direct information about the preceding changes in the target variables. Two parts of the inputs, describing the changes in the target variables from two different perspectives, can be obtained as follows:

$$\mathbf{D}^t = \mathrm{cat}\left(\mathbf{Y}^{t-1}, \Phi\left(\mathbf{Y}^{[1,T]}\right)^{t-1}\right) \tag{1}$$

where $\mathbf{D}^t$ is the input of the decoder at time point t, $\mathbf{Y}^{t-1}$ is the target variables at time point $t-1$, and $\Phi\left(\mathbf{Y}^{[1,T]}\right)^{t-1}$ represents the set of CTIs at time point $t-1$ obtained by using Granger causality analysis which contribute to the change in the target variables among T timestamps.

2.2 Spatial causal attention

Capturing long-term dependencies accurately between sequences is critical to improving the effectiveness of time series prediction. The self-attention mechanism can calculate the attention weights based on the similarity between m queries $\mathbf{Q} \in R^{m\times d}$ and m keys $\mathbf{K} \in R^{m\times d}$ using the dot-product operation in d-dimensional space and then obtain weighted values $\mathbf{V} \in R^{m\times d}$ as the attended representation.

The classical Transformer configured with embedding sub-layers was initially applied in natural language processing. However, time series variables carry genuine physical meanings, rendering the embedding sub-layers unnecessary. When adopting slow-changing temporal variables as the direct input of the attention mechanism, the similar dynamic information among adjacent time stamps and the

correlated variables can significantly affect the predictive performance, which may introduce redundant information and neglect other essential features.

The spatial and temporal dimensions of the time slices in multi-head dot-product attention can be regarded as hyperparameters, which can be determined through grid search or variable selection before model training. It is expected to have redundant settings to achieve better model performance. However, redundant settings can lead to redundant spatial information resulting from related and coupled variables. Therefore, attention will be paid to prioritizing the related sets of variables and assigning higher weights to them while assigning lower weights to other variables, leading to an inappropriately sparse weight matrix.

To reduce the correlation between variables and allocate equal attention to spatial dimensions representing different information, the spatial causal attention (SCA) projects the spatial dimensions of the query and key using mutually orthogonal projection vectors. After being projected, each spatial dimension represents a particular subspace of the original variables' features independently, which effectively reduces the redundancy and duplicate information between the variables and allows these approximately uncorrelated features to be regarded as causal features. SCA can be calculated with the following equation:

$$\boldsymbol{\alpha}^{i}_{\mathrm{CausalPos}} = \mathrm{softmax}\left(\frac{(\mathbf{QP}_i)\,(\mathbf{KP}_i)^T}{\sqrt{d'}}\right)\mathbf{V} \qquad (2)$$

where $\mathbf{P}_i \in R^{d\times d'}$ is the i^{th} spatial causal projection matrix and d' is the dimension of spatial projection.

2.3 Temporal causal attention

Long-term continuous and short-term rhythmic changes are prominent characteristics in temporal variables, so queries and keys at different time steps often show gradual variations, implying a similar attention weight distribution among different time steps. The lack of distinctiveness in the weighted values at different time steps is not beneficial for predicting multi-step target variables.

To address this problem, temporal causal attention (TCA) utilizes mutually orthogonal projection vectors to project the temporal dimensions of the key and the value, extracting causal features in the linear combinations of the original time slices that describe the prominent patterns of temporal changes within a pre-defined time interval. Through the semi-orthogonal transformation of the temporal dimension, the extracted temporal causal features exhibit an approximate uncorrelated relationship. TCA can be defined by:

$$\boldsymbol{\alpha}^{j}_{\mathrm{CausalTem}} = \mathrm{softmax}\left(\frac{\mathbf{Q}\left(\mathbf{M}_j^T\mathbf{K}\right)^T}{\sqrt{d}}\right)\left(\mathbf{M}_j^T\mathbf{V}\right) \qquad (3)$$

where $\mathbf{M}_j \in R^{m\times m'}$ is the j^{th} temporal causal projection matrix and m' is the dimension of temporal projection.

2.4 Multi-head spatial-temporal causal attention

Spatial-temporal causal attention (STCA) combines SCA and TCA by concatenating both to extract the spatial and temporal causal features from spatial and temporal dimensions separately. To fully extract causal features from different representation subspaces, multi-head STCA performs parallel computations of h individual STCA functions and concatenates the computed results before applying a linear projection for outputs, ultimately yielding the final result which can be obtained by:

$$\begin{aligned}\boldsymbol{\alpha}_{\mathrm{MultiCausal}} &= \mathrm{Concat}\left(\mathrm{head}_1,\cdots,\mathrm{head}_h\right)\mathbf{W}_o \\ \mathrm{head}_i &= \mathrm{Concat}\left(\alpha^{i}_{\mathrm{CausalPos}},\alpha^{i}_{\mathrm{CausalTem}}\right)\end{aligned} \qquad (4)$$

where $\mathbf{W}_o \in R^{2hd\times d}$ is the output matrix projecting the outputs of multiple STCA onto the feature dimension d to facilitate the stacking of other attention modules.

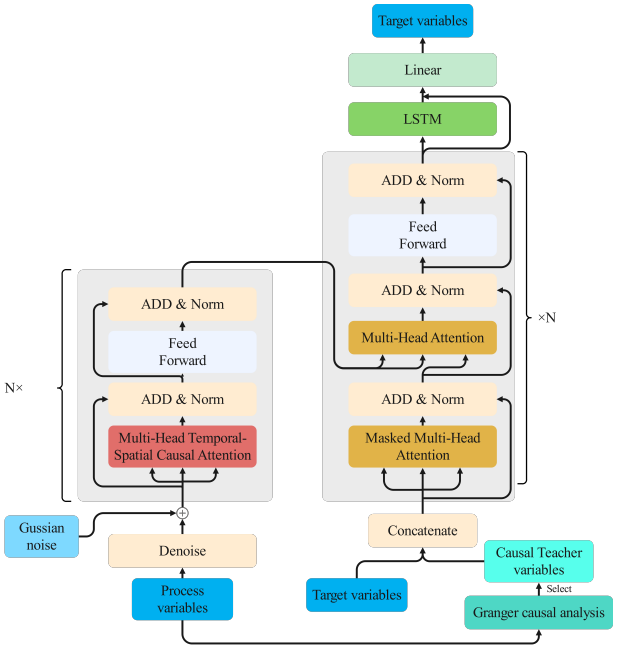


Fig. 1. The architecture of the CT

2.5 Causal-Transformer

The CT stacks multiple encoder and decoder blocks to capture long-term dependencies between slowly changing temporal variables. Fig. 1 shows the architecture of the CT which can be summarized as follows:

- The encoder block consists of a Multi-head STCA sub-layer to extract the temporal and spatial causal features and a position-wise fully connected feed-forward sub-layer, both adopting the residual structure and Batch normalization.
- The decoder block consists of a Multi-head attention sub-layer and a position-wise fully connected feed-forward sub-layer, with each sub-layer also adopting the residual structure and Batch normalization. Because CTIs can accelerate model convergence during training and provide stable guidance for predicting during testing based on the causal variables, the first decoder block takes the output of the top encoder block, historical values of the target variable, and CTIs as inputs.
- A long short-term memory (LSTM) sub-layer follows the top decoder block to capture the decoder block's long-term dynamic features of outputs. After that, a fully connected layer is added for the final prediction.

2.6 Loss function

On the one hand, to ensure predictive effectiveness, the CT uses mean square error (MSE) to penalize the deviation of model outputs from the actual values called prediction error loss. On the other hand, to ensure that spatial and temporal causal projection matrices represent orthogonal mappings of the spatial and temporal dimensions and to guarantee that the projection vector magnitude does not influence the feature projection components, the CT employs the Frobenius norm to constrain the unit length of projection vectors and the column orthogonality of spatial and the temporal causal projection matrices. Therefore, the loss function of the CT includes prediction error loss loss_{pre}, spatial causal projection loss loss_{pos}, and spatial causal projection loss loss_{tem}, which can be written as:

$$\begin{aligned}\text{Loss}_{\text{causal}} = &\underbrace{\text{MSE}\left(y,\hat{y}\right)}_{\text{loss}_{\text{pre}}} + \underbrace{\sum_{i=1}^{l}\sum_{j=1}^{h}\left\|\mathbf{P}_{ij}^{T}\mathbf{P}_{ij}-\mathbf{I}_{d'\times d'}\right\|_{F}}_{\text{loss}_{\text{pos}}} \\ &+ \underbrace{\sum_{i=1}^{l}\sum_{j=1}^{h}\left\|\mathbf{M}_{ij}^{T}\mathbf{M}_{ij}-\mathbf{I}_{m'\times m'}\right\|_{F}}_{\text{loss}_{\text{tem}}}\end{aligned} \tag{5}$$

where y and $\hat{y}$ are ground truth and prediction of target variables separately, l denotes the number of stacked encoder blocks, and $\mathbf{P}_{ij}$ and $\mathbf{M}_{ij}$ represent the spatial causal projection matrix and the temporal causal projection matrix corresponding to the j^{th} head in the i^{th} encoder block.

3. CASE STUDY

3.1 Experiment setup

The superiority of the CT has been validated in a real-world diesel refining process. As shown in Fig. 2, the feed in the form of vacuum gas oil after hydrotreating enters the V-1 vessel, where, after mixing with the recycle stream, it is fed into the R-1 hydrocracking reactor. The reaction mass from R-1 is mainly fed from a mixture of naphtha and diesel fractions into the product tank V-8. Light naphtha is distilled as the upper product of column C-1. The bottom product of column C-1 enters heater H-1 and then into column C-2 to separate the target diesel fraction from it. Part of the high-boiling C-2 product (residue) is recycled back to V-1. The presence of the liquid level in the bottom part of the C-2, recorded as variable LIC-C2 with a sampling interval of 1 minute, makes it possible to organize a stable flow of the recycle and prevent fluctuations in the R-1 reactor, thus regarded as the most critical variable for monitoring and prediction. The dataset originating from this diesel refining process consists of 1200 samples, with the first 1000 points utilized as the train set and the remaining data points employed as the test set.

To compare the experimental effects of different models, this paper trained LSTM, gated recurrent unit (GRU) network, Transformer (Vaswani et al., 2017), Informer (Zhou et al., 2021) and CT combining with LSTM, one-dimensional convolution neural network (1D-CNN), and fully connected neural network within five time steps and

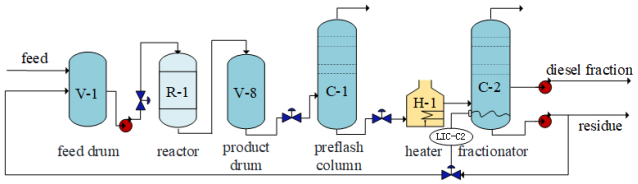


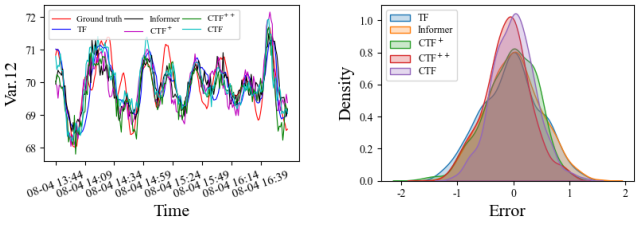
Fig. 2. The schematic diagram of a diesel refining process

chose MSE and R^2 as evaluation metrics. Each model determined its optimal hyperparameters through grid search.

3.2 Result analysis

Prediction performance Table 1 and Fig. 3 show the five-step prediction results of the six models, LSTM, GRU, Transformer, CT^{+},CT^{++}, and CT, where the CT^{+} and the CT^{++} replace the LSTM sub-layer in the CT with the fully connected layers and 1D-CNN separately. The following conclusions can be drawn:

- It is common for LSTM and GRU models to suffer from accumulating prediction errors in multi-step time series prediction problems.
- Compared to the classic Transformer and Informer, the CT shows a significant improvement on the refinery dataset due to the introduction of causality, which means a reduction of 46.0% in MSE towards the classic Transformer and a reduction of 30.4% in MSE towards the Informer.
- Compared to the CT^{+}, the CT^{++} utilizes 1D-CNN to extract features of different granularities in the spatial dimension, resulting in a 14.9% reduction in MSE and an 8.6% increase in R^2. The CT employs LSTM as a multi-step prediction regressor, enabling the exploration of dynamic characteristics among predicted results at different time steps and achieving a 21.4% reduction in MSE and a 9.6% increase in R^2 compared to the CT^{++}.



(a) The fifth time step (b) Error distribution

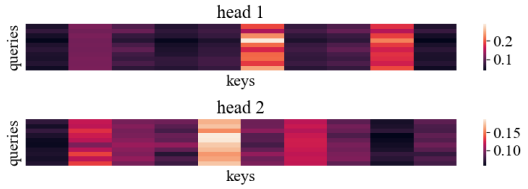
Fig. 3. Results of five-step prediction

Attention weights The Transformer computes attention weights on the original slow-changing time series data, leading to significant similarity in the attention weights between different queries. Fig. 4 shows that although different heads in a Transformer can extract information from different representation subspaces, the high similarity of attention weights at different positions within a head is disadvantageous for extracting different crucial features for multi-step prediction. However, multi-head STCA compresses features of spatial and temporal dimensions before computing attention weights, which helps to condense

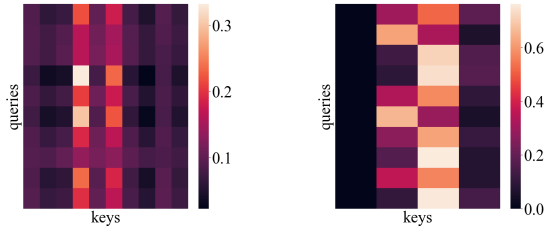
Table 1. Performance comparison for different methods on a refining case

Model	Training set		Test set	
	MSE	R^2	MSE	R^2
LSTM	2.3922×10^{-5}	0.99997	0.28655	0.64794
GRU	$\mathbf{1.1689 \times 10^{-5}}$	**0.99998**	0.33672	0.58630
Transformer	0.13074	0.82033	0.36786	0.54804
Informer	0.096535	0.86717	0.28532	0.62746
CT^{+}	0.047426	0.93475	0.29683	0.63531
CT^{++}	0.024534	0.96632	0.25256	0.68970
CT	0.017637	0.97585	**0.19852**	**0.75609**

similar information. By computing attention weights in nearly orthogonal spatial and temporal dimensions, it alleviates the problem of excessive similarity among queries and keys, significantly increasing discriminability among different queries.



(a) Attention weights in a Transformer



(b) Attention weights of SCA (c) Attention weights of TCA

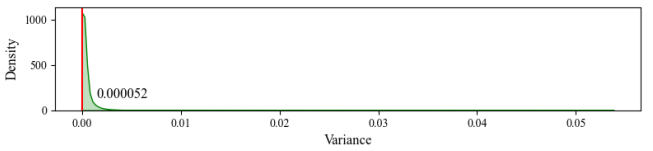
Fig. 4. Attention weights in different models

Fig. 5 displays the kernel density estimation of the variance distribution of attention weights in different attention-based models. The red line indicates the maximum variance in the kernel density curve. It can be seen that the variance of attention weights in models using causal attention is significantly more significant than that in the Transformer. Furthermore, the variance of attention weights in the TCA is generally greater than that of SCA on the refining dataset, which suggests that the temporal causal attention is more capable of generating more considerable variations in attention weights, helping the model better capture dynamic changes along the time dimension.

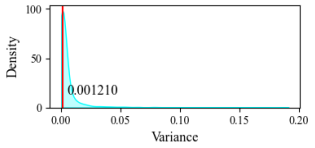
3.3 Ablation

Different components of the CT were successively removed to verify the separate effect, and hyperparameter selection and training were then carried out for each modified model. The prediction performance of the variants on the training and testing sets is shown in Table 2 from which the following conclusions can be drawn:

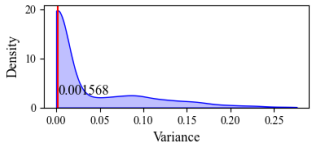
- Residual connections in the LSTM sub-layer can help the model generalize better and reduce overfitting.



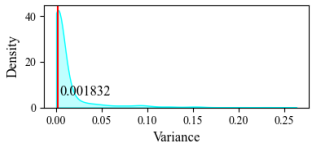
(a) Transformer



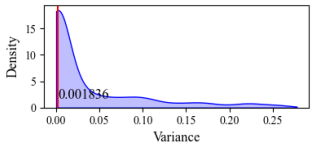
(b) SCA of CT^{+}



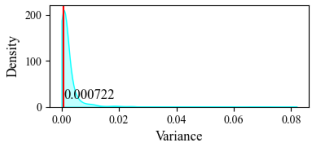
(c) TCA of CT^{+}



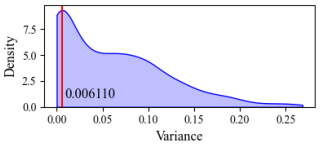
(d) SCA of CT^{++}



(e) TCA of CT^{++}



(f) SCA of CT



(g) TCA of CT

Fig. 5. Kernel density estimation of variance distribution of attention weights

- Without using CTIs the model's performance decreases substantially but still outperforms the Transformer. But using all available variables as the inputs of the decoder block does not improve the performance of the model further. The selection of causal variables has a marked impact on the model, and a reasonable combination of causal variables enables the model to achieve better performance.
- Not using causal attention leads to an apparent decrease in performance. However, the stable assistance of CTIs makes the model's performance better than that of Transformers.
- Considering that the CT directly takes process variables as input and the time series data itself carries relative positional temporal information, positional encodings are not adopted in the CT. Besides that, the result shows an increase of 17.9% in MSE and

Table 2. Variations on the CT

Model	Training set		Test set	
	MSE	R^2	MSE	R^2
w/o Residual connection	0.014656	0.97989	$0.20940_{+0.011}$	$0.74272_{-0.013}$
w/ All variables	0.018785	0.97417	$0.26739_{+0.069}$	$0.67148_{-0.085}$
w/o All CTIs	0.10637	0.85454	$0.32140_{+0.12}$	$0.60512_{-0.15}$
w/o STCA	0.020507	0.97196	$0.23464_{+0.036}$	$0.71172_{-0.044}$
w/ Positional encodings	0.019530	0.97325	$0.23400_{+0.035}$	$0.70513_{-0.051}$

a decrease of 6.7% in R^2 with positional encoding, indicating that positional encoding does not improve the model's performance on the refining dataset.

4. CONCLUSION

To alleviate the redundant information among queries and keys in predicting slowly varying time series data using Transformer-based models, this paper proposes the multi-head STCA to perform semi-orthogonal projection on queries and keys to extract independent spatial and temporal causal features. The CT regards CTIs as part of the decoder's input, which provides fuller and more stable causal assistance when facing the accumulation of prediction deviations. On the refinery dataset, the CT achieved the best performance with an MSE of 0.19852 on the test set. Future work will focus on the integration of causal analysis and graph neural networks.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (61873142, 62111530057) and the Federal State Budgetary of the Institute of Automation and Control Processes FEB RAS under topic No. FWFW-2021-0003.

REFERENCES

Chen, J., Li, K., Rong, H., Bilal, K., Li, K., and Philip, S.Y. (2019). A periodicity-based parallel time series prediction algorithm in cloud computing environments. *Information Sciences*, 496, 506–537.

Dong, P., Lian, J., and Zhang, Y. (2019). A novel data-driven approach for tropical cyclone tracks prediction based on Granger causality and GRU. In *2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 70–75. IEEE.

Geng, Z., Chen, Z., Meng, Q., and Han, Y. (2022). Novel Transformer Based on Gated Convolutional Neural Network for Dynamic Soft Sensor Modeling of Industrial Processes. *IEEE Transactions on Industrial Informatics*, 18(3), 1521–1529. doi:10.1109/TII.2021.3086798.

Han, J., Zhang, X.P., and Wang, F. (2016). Gaussian process regression stochastic volatility model for financial time series. *IEEE Journal of Selected Topics in Signal Processing*, 10(6), 1015–1028.

Kashpruk, N., Piskor-Ignatowicz, C., and Baranowski, J. (2023). Time Series Prediction in Industry 4.0: A Comprehensive Review and Prospects for Future Advancements. *Applied Sciences*, 13(22).

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X., and Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, 32.

Li, X., Feng, S., Hou, N., Li, H., Zhang, S., Jian, Z., and Zi, Q. (2022). Applications of Kalman Filtering in Time Series Prediction. In *International Conference on Intelligent Robotics and Applications*, 520–531. Springer.

Lim, B. and Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209.

Liu, C., Hoi, S.C., Zhao, P., and Sun, J. (2016). Online arima algorithms for time series prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Liu, M., Wang, W., Hu, X., Fu, Y., Xu, F., and Miao, X. (2023). Multivariate long-time series traffic passenger flow prediction using causal convolutional sparse self-attention MTS-Informer. *Neural Computing and Applications*, 1–17.

Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., and Sheth, A.P. (2018). Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks*, 4(3), 161–175.

Pearl, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 3.

Sapankevych, N.I. and Sankar, R. (2009). Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, 4(2), 24–38.

Shen, L. and Wang, Y. (2022). TCCT: Tightly-coupled convolutional transformer on time series forecasting. *Neurocomputing*, 480, 131–145.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Yu, F., Xiong, Q., Cao, L., and Yang, F. (2022). Stable soft sensor modeling based on causality analysis. *Control Engineering Practice*, 122, 105109.

Yuan, X., Li, L., Shardt, Y.A.W., Wang, Y., and Yang, C. (2021). Deep Learning With Spatiotemporal Attention-Based LSTM for Industrial Soft Sensor Model Development. *IEEE Transactions on Industrial Electronics*, 68(5), 4404–4414.

Zhang, Y. and Yan, J. (2022). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11106–11115.