A Details of Datasets

Electricity The UCI electricity load diagrams dataset [Mitrea *et al.*, 2009] contains 370 customer power consumption per unit time. There is no missing value in this dataset, recording the power consumption per 15 minutes (KWH) from 2011 to 2014. Each column of the data represents a customer (370 columns in total), each row represents a quarter (140,256 rows in total), and all time labels are subject to Portuguese time.

Air-Quality The Air-Quality dataset [UCI, 2006] contains 9,358 instances of hourly averaged responses from an array of 5 metal chemical sensors embedded in Air quality Device. Data was recorded from March 2004 to February 2005 (one year). The device was located on the field in a significantly polluted area.

Industry data The data of Hangseng Stock Composite Index (HSCI) and another eleven industry stock indices are obtained from the Wind platform. Eleven industry stock indices include consumer good manufacturing, consumer service, energy, finance, industry, information technology, integrated industry, raw material, real estate, and utilities. The dataset covers the time period from September 2006 up to September 2019.

Datasets	T	D	L
Electricity	26,304	321	1 hour
Air-Quality	9,358	12	1 hour
Industry Stock Composite Index	3,205	12	1 day

Table 3: Dataset statistics, where T is length of time series or data size, D is the number of variables, and L is the sample rate. The three datasets are representative (from easy to difficult). The sample rate ranges from hour to day, and the dimension is from 12 to 321.

B Metric Description

We choose two widely used metrics to measure the performance on multivariate time series forecasting datasets. The first one is the root relative squared error (RSE), which is the scaled version of the widely used Root Mean Square Error (RMSE) to remove the influence of data scale:

$$RSE = \frac{\sqrt{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n}(y_{t,i} - \hat{y_{t,i}})^2}}{\sqrt{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n}(\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_1,1:n}})^2}} \tag{9}$$

The second metric is the empirical correlation coefficient (CORR), and the CORR is the measure of correlation between actual and forecast variables in time series.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_{t,i} - \overline{y_{t_0:\hat{t}_{1,i}}})(\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_{1,i}}})}{\sqrt{\sum_{t=t_0}^{t_1}(y_{t,i} - \overline{y_{t_0:t_{1:i}}})^2 \sum_{t=t_0}^{t_1}(\hat{y}_{t,i} - \overline{\hat{y}_{t_0:t_{1:i}}})^2}} \tag{10}$$

where y and $\hat{y}$ are the ground truth and the predicted value, respectively. $\overline{y}$ denotes the mean of set y values within a given set (e.g., testing set). For RSE, the lower the better, whereas for CORR, the higher the better. These datasets and the metrics of RSE and CORR have been widely used in many papers about time series forecasting.

C Detailed Experimental Settings

C.1 Base module

The four state-of-art deep learning methods for comparison are CNN, GRU, LSTNet and Self-Attention encoder, as detailed below:

- *CNN*: Convolutional nerual network [Goodfellow *et al.*, 2016] was designed to ensure only past information for forecasting. We use 7-layer CNN to do time series forecasting.
- *GRU + Attention*: GRU [Chung *et al.*, 2014] had been used in time series forecasting and combined with attention to improve interpretability.
- *LSTNet*: LSTNet [Lai *et al.*, 2018] is a model using CNN and RNN for multivariate time series forecasting. The architecture uses convolutional network and the Recurrent Neural Network (RNN) to extract short-term local dependency patterns among variables and to discover long-term patterns for time series trends.
- *Self-Attention encoder*: Transformer-based [Vaswani *et al.*, 2017] architecture has been modified for forecasting. We design a variant of Transformer with encoder-only structure which consists of L blocks of multi-head Self-Attention layers and position-wise feed forward layers.

C.2 Hyperparameters

Since the model structure is universal for all methods, we adjust the same optimal hyperparameters on the training data. Firstly, we use the Adam algorithm for the optimization with learning rate 10^{-4} and weigth decay 10^{-3}. For data preprocessing, we scale the data into the range $[0, 1]$ by batch normalization to avoid extreme values and improve the computation stability. We set $\lambda_1 = 10^{-3}$ and $\lambda_2 = 10^{-3}$ in both L_1 and L_2. We select the batch size according to the size of each dataset (either 64 or 128).

D Forecasting Visualization and Interpretation Features on More Datasets

Because the Self-Attention encoder obtains the best performance, we evaluate the forecasting results of Self-Attention encoder visually in series saliency on all the three datasets. As shown in Fig.s 7,8,9, our method gives accurate forecasts. The proposed model clearly yields better forecasts around the flat line after the peak and in the valley.

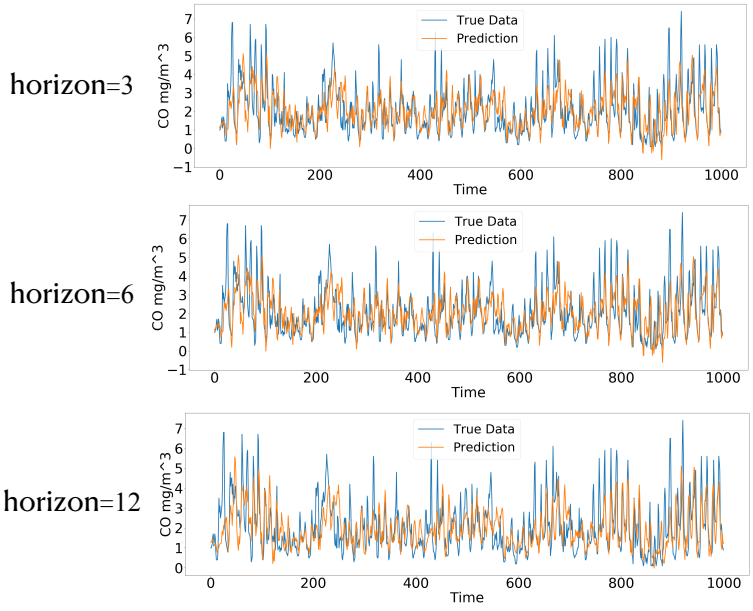

Figure 7: Prediction results for self-attention encoder in series saliency on air quality with horizon $\tau = \{3, 6, 12\}$ and window size $T = 64$. The feature is the true hourly averaged concentration CO in mg/m^3

Because the Self-Attention encoder obtains the best performance, we visualize the series saliency on other datasets, including the air quality and industry stock composite index, for horizon $\tau = 6$. As shown in Fig. 10 and Fig. 11, we could analyze the series saliency and obtain the corresponding feature importance.

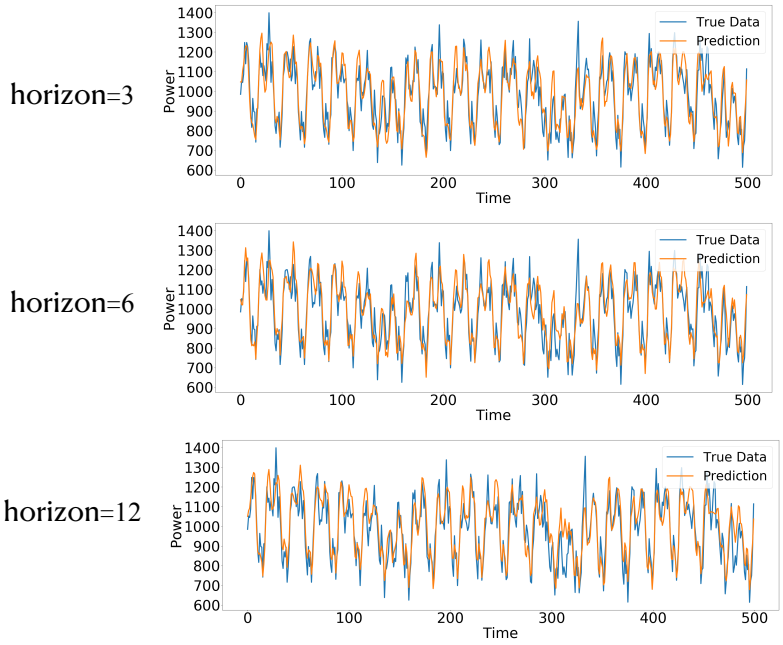

Figure 8: Prediction results for Self-Attention encoder in series saliency on electricity with horizon $\tau = \{3, 6, 12\}$ and window size $T = 168$. The feature is power consumption of No.7 powerplant. The model learned the periodicity of electricity data.

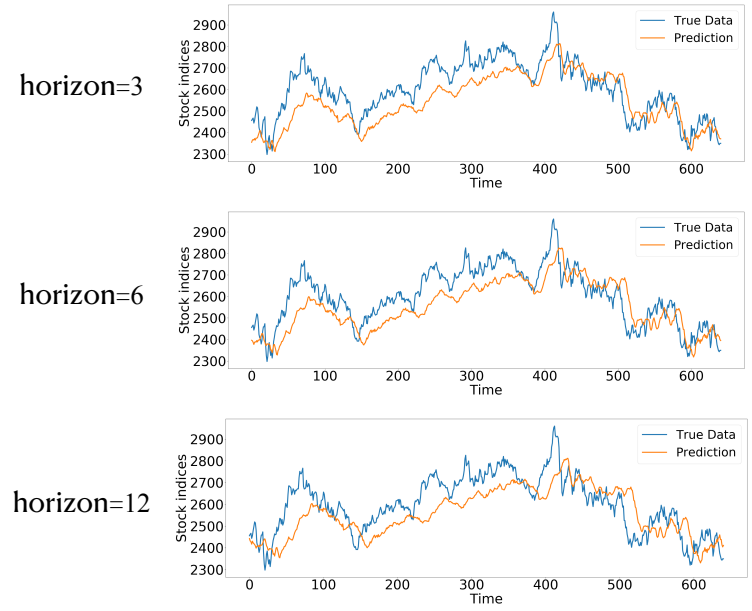

Figure 9: Prediction results for Self-Attention encoder in series saliency on industry stock indicies with $\tau = \{3, 6, 12\}$ and window size $T = 168$. The feature is No.1 stock indicies of Hangseng Stock Composite Index.

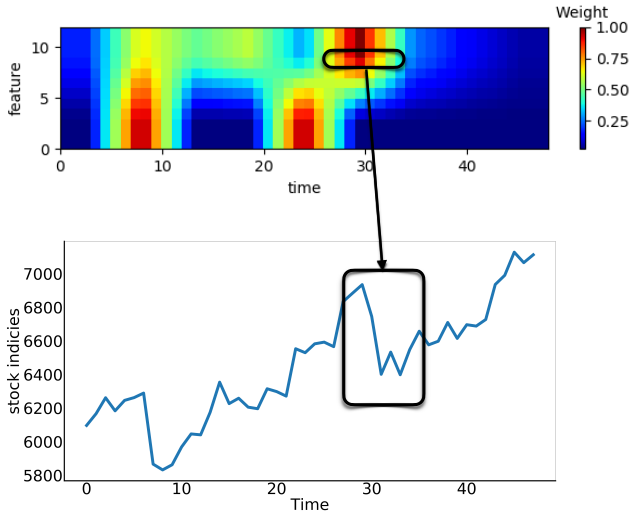

Figure 10: Series saliency for Self-Attention encoder on industry stock indicies with horizon $\tau = 6$ and window size $T = 64$. The highlighted area corresponds to the dramatic change in the industry stock indices. The phenomenon shows that the stock indices influences the forecasting greatly.

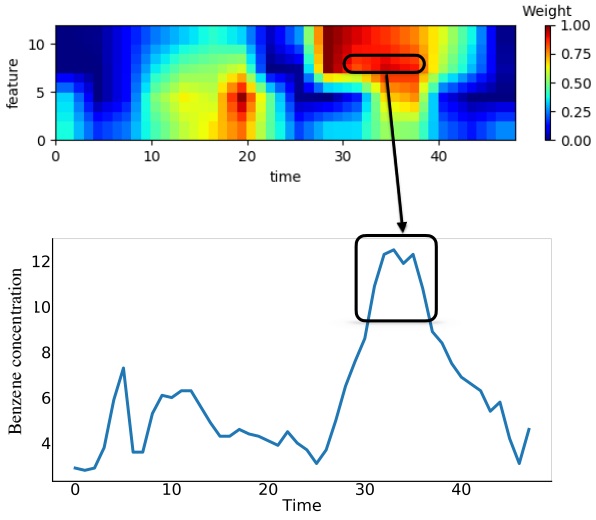

Figure 11: Series saliency for Self-Attention encoder on air quality with horizon $\tau = 6$ and window size $T = 64$. The highlighted area corresponds that benzene increases sharply in air quality. As you known, benzene is harmful, which impacts air quality greatly.

E Empirical results

The full set of results (including RSE and CORR) is shown in Table 4 and Table 5. We ran the experiments for 10 times and present the full results (mean and standard deviations).

	Air Quality			Industry			Electricity		
Methods	$\tau = 3$	$\tau = 6$	$\tau = 12$	$\tau = 3$	$\tau = 6$	$\tau = 12$	$\tau = 3$	$\tau = 6$	$\tau = 12$
CNN	0.775 ± 0.003	0.701 ± 0.001	0.636 ± 0.001	0.919 ± 0.022	0.919 ± 0.019	0.841 ± 0.008	0.883 ± 0.004	0.871 ± 0.002	0.866 ± 0.004
GRU	0.804 ± 0.003	0.712 ± 0.002	0.639 ± 0.003	0.953 ± 0.003	0.936 ± 0.013	0.904 ± 0.011	0.878 ± 0.001	0.877 ± 0.003	0.867 ± 0.002
LSTNet	0.777 ± 0.001	0.708 ± 0.004	0.623 ± 0.004	0.949 ± 0.004	0.934 ± 0.003	0.876 ± 0.011	0.922 ± 0.004	0.913 ± 0.002	0.906 ± 0.002
Self-Attention	0.813 ± 0.002	0.722 ± 0.003	0.643 ± 0.003	0.961 ± 0.002	0.942 ± 0.005	0.905 ± 0.009	0.919 ± 0.007	0.907 ± 0.001	0.902 ± 0.003
CNN w/ Type1	0.765 ± 0.002	0.712 ± 0.001	0.637 ± 0.003	0.923 ± 0.006	0.921 ± 0.002	0.835 ± 0.001	0.889 ± 0.002	0.882 ± 0.007	0.876 ± 0.001
GRU w/ Type1	0.807 ± 0.004	0.713 ± 0.005	0.642 ± 0.011	0.959 ± 0.009	0.926 ± 0.002	0.905 ± 0.012	0.883 ± 0.001	0.879 ± 0.006	0.869 ± 0.003
LSTNet w/ Type1	0.788 ± 0.001	0.683 ± 0.002	0.643 ± 0.002	0.952 ± 0.007	0.931 ± 0.001	0.879 ± 0.006	0.922 ± 0.001	0.912 ± 0.005	0.906 ± 0.008
SA w/ Type1	0.814 ± 0.001	0.717 ± 0.005	0.639 ± 0.004	0.963 ± 0.003	0.949 ± 0.003	0.907 ± 0.001	0.919 ± 0.007	0.908 ± 0.004	0.903 ± 0.003
CNN w/ Type2	0.764 ± 0.001	0.721 ± 0.002	0.639 ± 0.003	0.944 ± 0.005	0.924 ± 0.003	0.864 ± 0.007	0.891 ± 0.001	0.891 ± 0.003	0.882 ± 0.006
GRU w/ Type2	0.808 ± 0.004	0.710 ± 0.001	0.641 ± 0.003	0.957 ± 0.008	0.939 ± 0.007	0.911 ± 0.003	0.899 ± 0.002	0.883 ± 0.007	0.877 ± 0.005
LSTNet w/ Type2	0.791 ± 0.003	0.704 ± 0.001	0.645 ± 0.004	0.956 ± 0.007	0.934 ± 0.008	0.887 ± 0.002	0.924 ± 0.003	0.916 ± 0.003	0.904 ± 0.008
SA w/ Type2	0.814 ± 0.001	0.729 ± 0.006	0.646 ± 0.005	0.954 ± 0.002	0.953 ± 0.003	0.914 ± 0.002	0.926 ± 0.003	0.911 ± 0.001	0.903 ± 0.003
CNN w/ SS	0.779 ± 0.005	0.723 ± 0.009	0.641 ± 0.007	0.941 ± 0.006	0.927 ± 0.004	0.881 ± 0.001	0.898 ± 0.004	0.893 ± 0.002	0.892 ± 0.007
GRU w/ SS	0.809 ± 0.003	0.716 ± 0.012	0.649 ± 0.003	0.955 ± 0.001	0.935 ± 0.002	0.912 ± 0.003	0.905 ± 0.004	0.889 ± 0.008	0.878 ± 0.003
LSTNet w/ SS	0.794 ± 0.008	0.724 ± 0.002	0.641 ± 0.003	0.959 ± 0.004	0.938 ± 0.001	0.901 ± 0.002	$\mathbf{0.928} \pm 0.003$	$\mathbf{0.918} \pm 0.003$	0.907 ± 0.001
SA w/ SS	$\mathbf{0.819} \pm 0.003$	$\mathbf{0.732} \pm 0.009$	$\mathbf{0.658} \pm 0.001$	$\mathbf{0.965} \pm 0.003$	$\mathbf{0.955} \pm 0.016$	$\mathbf{0.916} \pm 0.004$	0.923 ± 0.003	0.915 ± 0.001	$\mathbf{0.911} \pm 0.002$

Table 4: Empirical CORR results of air quality, industry and electricity when horizon $\tau = \{3, 6, 12\}$. Best performance in boldface. We report the mean and standard deviations in ten runs. For CORR, the higher the better.

	Air Quality			Industry			Electricity		
Methods	$\tau = 3$	$\tau = 6$	$\tau = 12$	$\tau = 3$	$\tau = 6$	$\tau = 12$	$\tau = 3$	$\tau = 6$	$\tau = 12$
CNN	0.309 ± 0.001	0.362 ± 0.007	0.405 ± 0.002	0.162 ± 0.003	0.168 ± 0.006	0.208 ± 0.008	0.101 ± 0.004	0.105 ± 0.006	0.107 ± 0.004
GRU	0.312 ± 0.011	0.357 ± 0.012	0.397 ± 0.002	0.202 ± 0.006	0.192 ± 0.009	0.222 ± 0.001	0.119 ± 0.007	0.125 ± 0.005	0.132 ± 0.004
LSTNet	0.327 ± 0.003	0.362 ± 0.006	0.407 ± 0.008	0.181 ± 0.010	0.191 ± 0.005	0.228 ± 0.004	0.089 ± 0.001	0.098 ± 0.003	0.104 ± 0.006
Self-Attention	0.301 ± 0.002	0.353 ± 0.003	0.387 ± 0.008	0.147 ± 0.006	0.172 ± 0.007	0.186 ± 0.001	0.088 ± 0.008	0.095 ± 0.007	0.101 ± 0.010
CNN w/ Type1	0.303 ± 0.006	0.361 ± 0.002	0.401 ± 0.002	0.168 ± 0.005	0.161 ± 0.004	0.199 ± 0.005	0.099 ± 0.003	0.101 ± 0.005	0.103 ± 0.006
GRU w/ Type1	0.307 ± 0.009	0.342 ± 0.002	0.386 ± 0.005	0.185 ± 0.003	0.184 ± 0.004	0.214 ± 0.006	0.113 ± 0.009	0.119 ± 0.004	0.128 ± 0.009
LSTNet w/ Type1	0.314 ± 0.008	0.361 ± 0.008	0.399 ± 0.006	0.180 ± 0.007	0.189 ± 0.003	0.225 ± 0.002	0.088 ± 0.003	0.092 ± 0.005	0.099 ± 0.006
SA w/ Type1	0.298 ± 0.004	0.351 ± 0.002	0.366 ± 0.005	0.134 ± 0.006	0.169 ± 0.001	0.181 ± 0.002	0.086 ± 0.003	0.096 ± 0.005	0.097 ± 0.005
CNN w/ Type2	0.299 ± 0.003	0.355 ± 0.005	0.398 ± 0.004	0.158 ± 0.008	0.164 ± 0.006	0.196 ± 0.009	0.093 ± 0.003	0.099 ± 0.004	0.101 ± 0.001
GRU w/ Type2	0.301 ± 0.003	0.337 ± 0.002	0.356 ± 0.001	0.179 ± 0.004	0.187 ± 0.007	0.205 ± 0.006	0.109 ± 0.008	0.115 ± 0.003	0.119 ± 0.002
LSTNet w/ Type2	0.299 ± 0.007	0.356 ± 0.002	0.389 ± 0.001	0.178 ± 0.008	0.193 ± 0.005	0.222 ± 0.001	0.087 ± 0.011	0.093 ± 0.006	0.098 ± 0.005
SA w/ Type2	0.291 ± 0.001	0.348 ± 0.003	0.362 ± 0.006	0.125 ± 0.006	0.166 ± 0.001	0.178 ± 0.002	0.085 ± 0.002	0.091 ± 0.002	0.094 ± 0.001
CNN w/ SS	0.298 ± 0.007	0.351 ± 0.004	0.392 ± 0.001	0.153 ± 0.004	0.166 ± 0.002	0.191 ± 0.003	0.090 ± 0.002	0.103 ± 0.001	0.102 ± 0.004
GRU w/ SS	0.292 ± 0.004	$\mathbf{0.323} \pm 0.002$	$\mathbf{0.342} \pm 0.006$	0.177 ± 0.009	0.176 ± 0.007	0.199 ± 0.002	0.106 ± 0.008	0.112 ± 0.005	0.112 ± 0.009
LSTNet w/ SS	0.297 ± 0.009	0.349 ± 0.004	0.387 ± 0.000	0.173 ± 0.001	0.179 ± 0.002	0.218 ± 0.005	0.086 ± 0.004	0.094 ± 0.008	0.094 ± 0.006
SA w/ SS	$\mathbf{0.288} \pm 0.003$	0.341 ± 0.003	0.351 ± 0.003	$\mathbf{0.119} \pm 0.003$	$\mathbf{0.159} \pm 0.003$	$\mathbf{0.174} \pm 0.003$	$\mathbf{0.081} \pm 0.002$	$\mathbf{0.089} \pm 0.001$	$\mathbf{0.091} \pm 0.002$

Table 5: Empirical RSE results of air quality, industry and electricity when horizon $\tau = \{3, 6, 12\}$. Best performance in boldface. We report the mean and standard deviations in ten runs. For RSE, the lower the better.