Published as a conference paper at ICLR 2024

iTransformer: Inverted Transformers Are Effective for Time Series Forecasting

Yong Liu,* Tengge Hu,* Haoran Zhang,* Haixu Wu, Shiyu Wang[§], Lintao Ma[§], Mingsheng Long[✉]
School of Software, BNRist, Tsinghua University, Beijing 100084, China
[§]Ant Group, Hangzhou, China
{liuyong21,htg21,z-hr20,whx20}@mails.tsinghua.edu.cn
{weiming.wsy,lintao.mlt}@antgroup.com, mingsheng@tsinghua.edu.cn


Yong Liu


Tengge Hu


Haoran Zhang


Haixu Wu


Jianmin Wang


Mingsheng Long

Time Series Forecasting

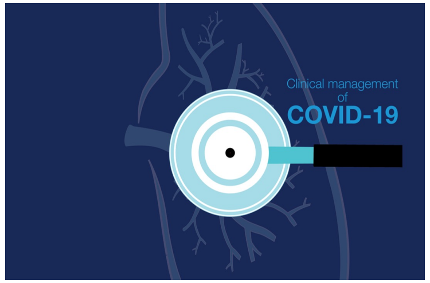
Wide Applications

Energy Consumption

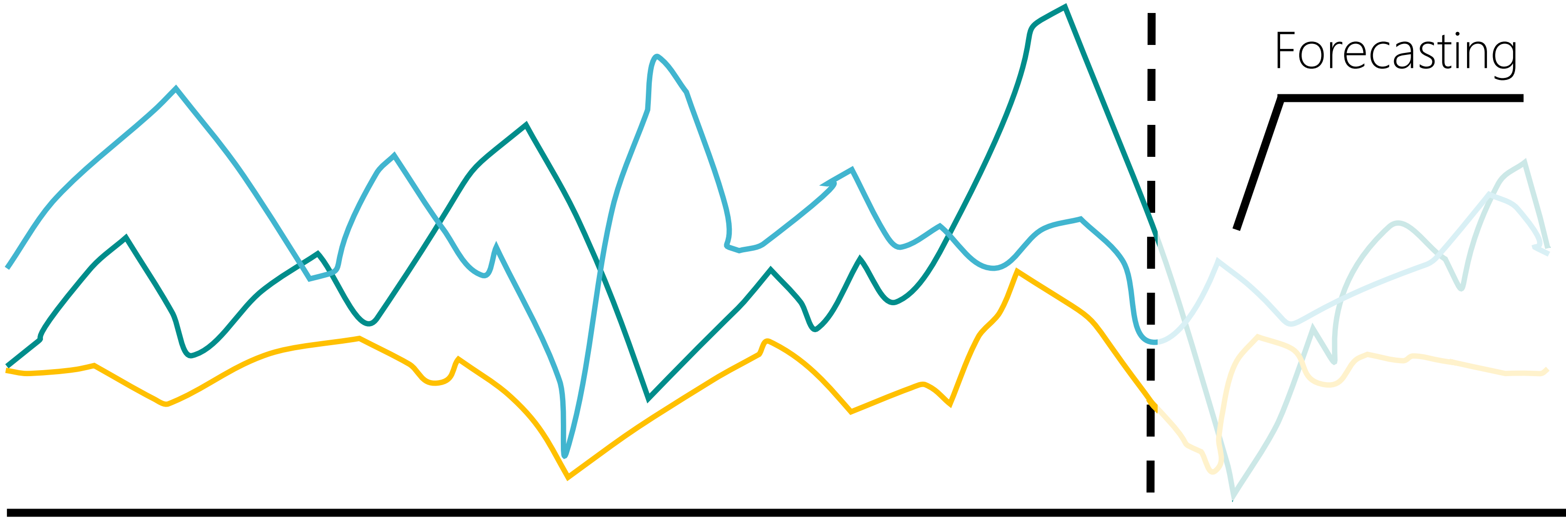
Traffic Flow

Economic Changes

Weather Variations



Disease Estimations



Challenges

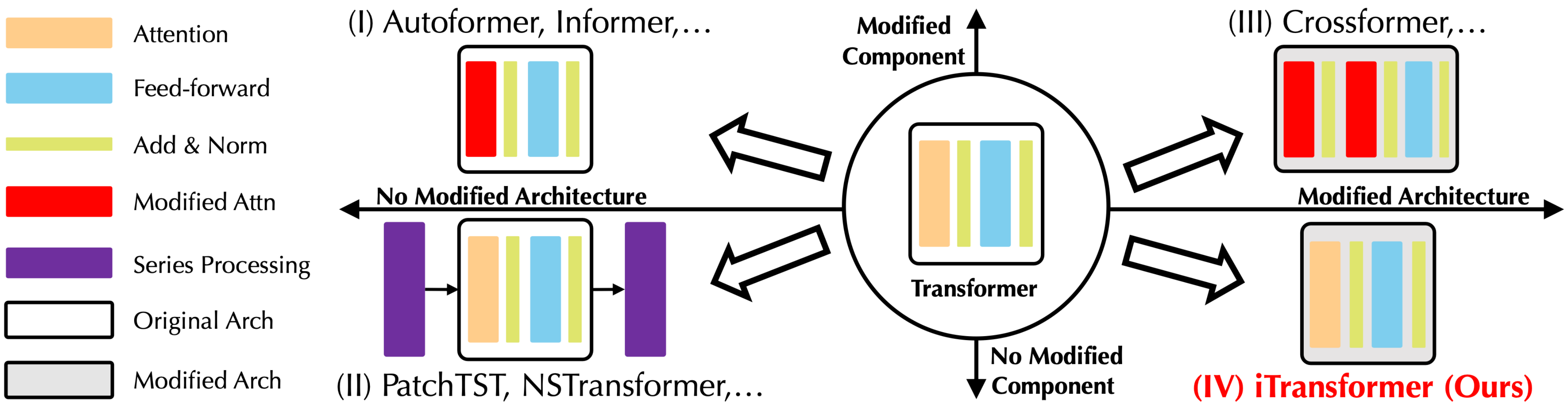
- Complicated Variations
- Numerous Variates

Past Observations

Future Time Series

Transformer-based Forecaster

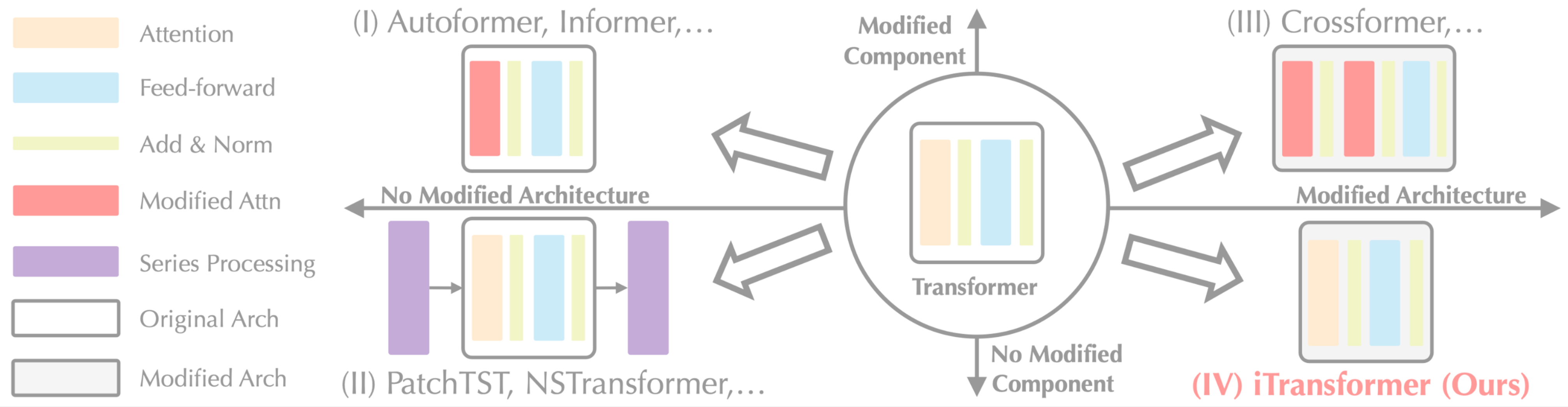
- Emergence of Transformers in TSF



- Passionate modifications!

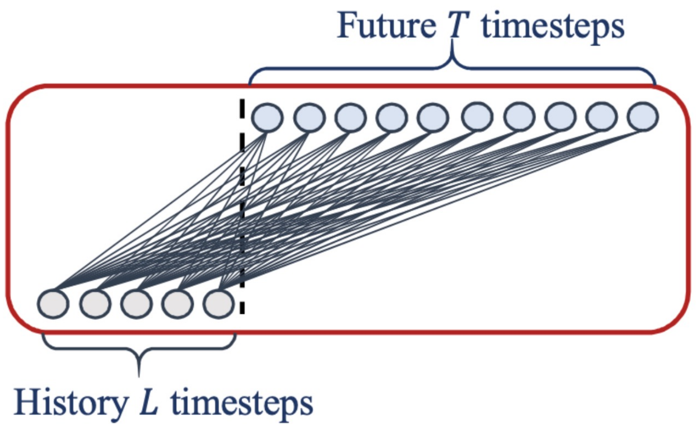
Transformer-based Forecaster

- Emergence of Transformers in TSF



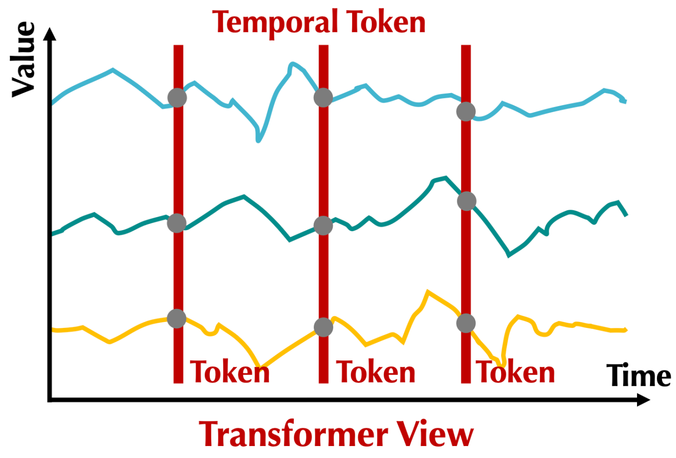
- Passionate modifications!

- Linear models beat Transformers?
  - ARIMA, Holt-Winter …
  - DLinear, RLinear …



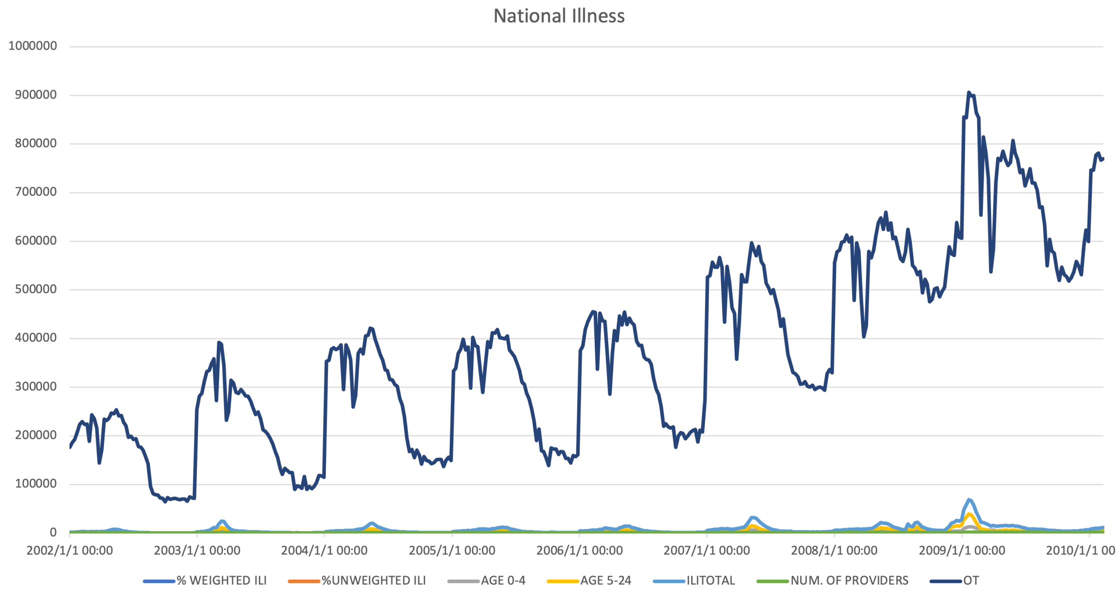
Zeng et al. Are Transformers Effective for Time Series Forecasting? *AAAI 2023.*

Time Series Tokens in Transformer

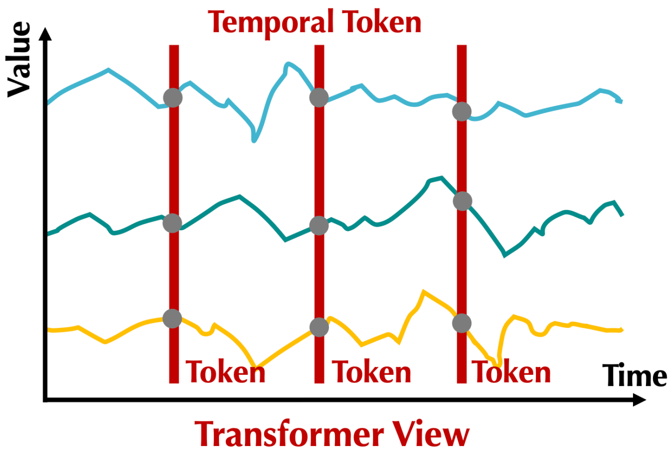


Underlying Risks of Tokenization

- Excessively receptive field
- Inconsistent scale and distribution

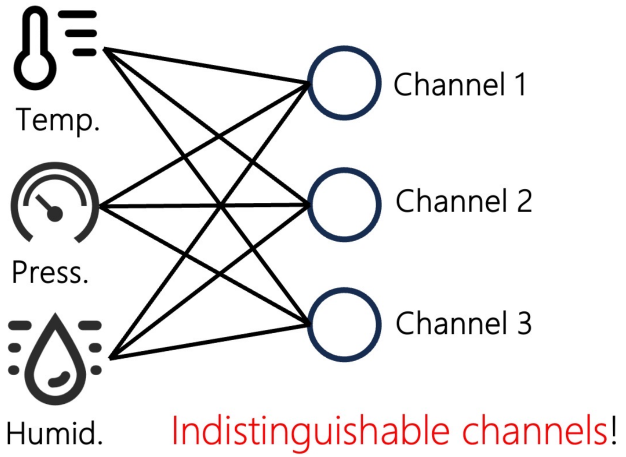


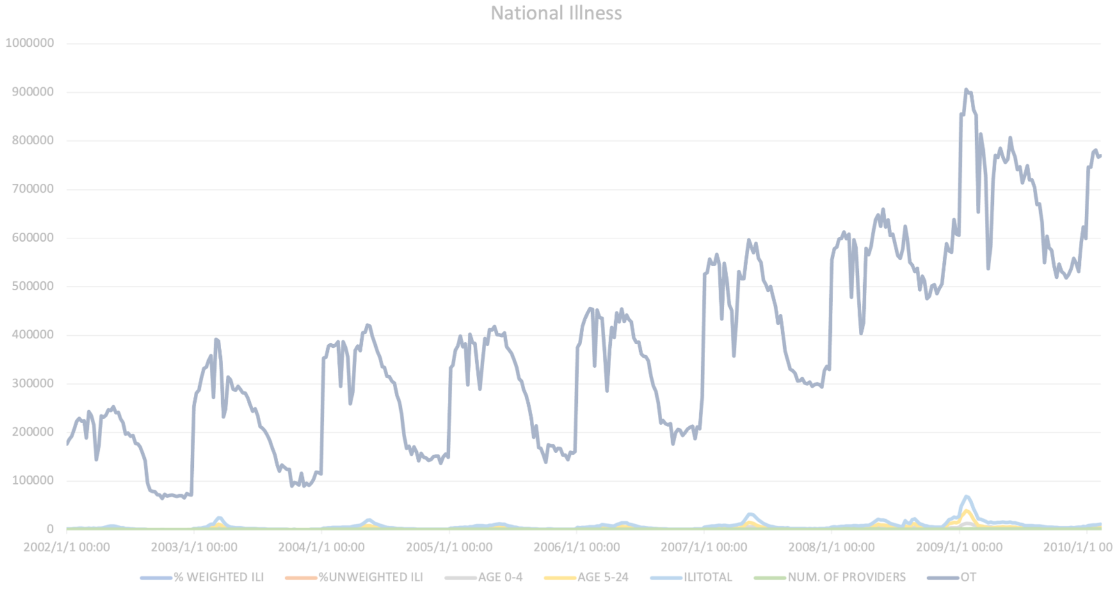
Time Series Tokens in Transformer

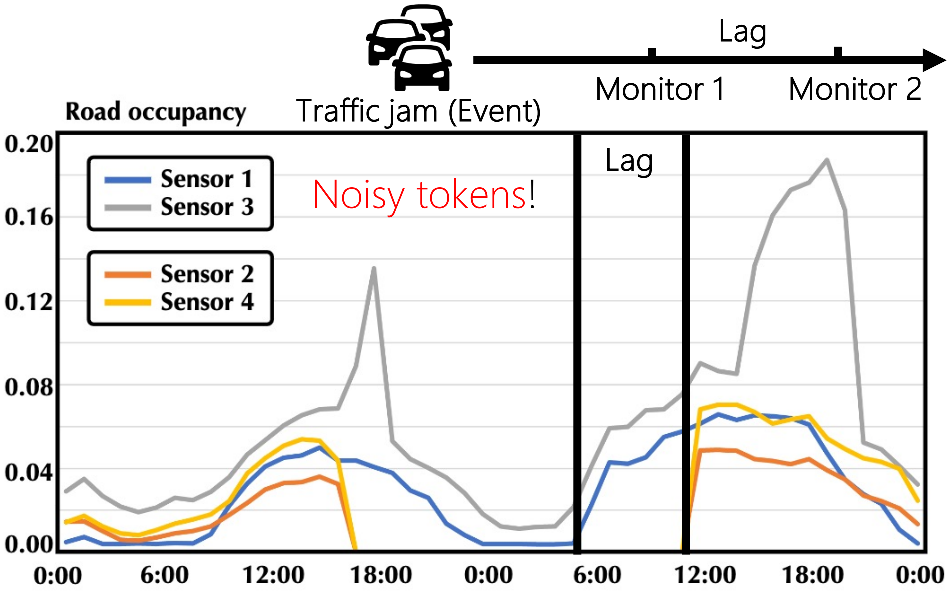


Underlying Risks of Tokenization

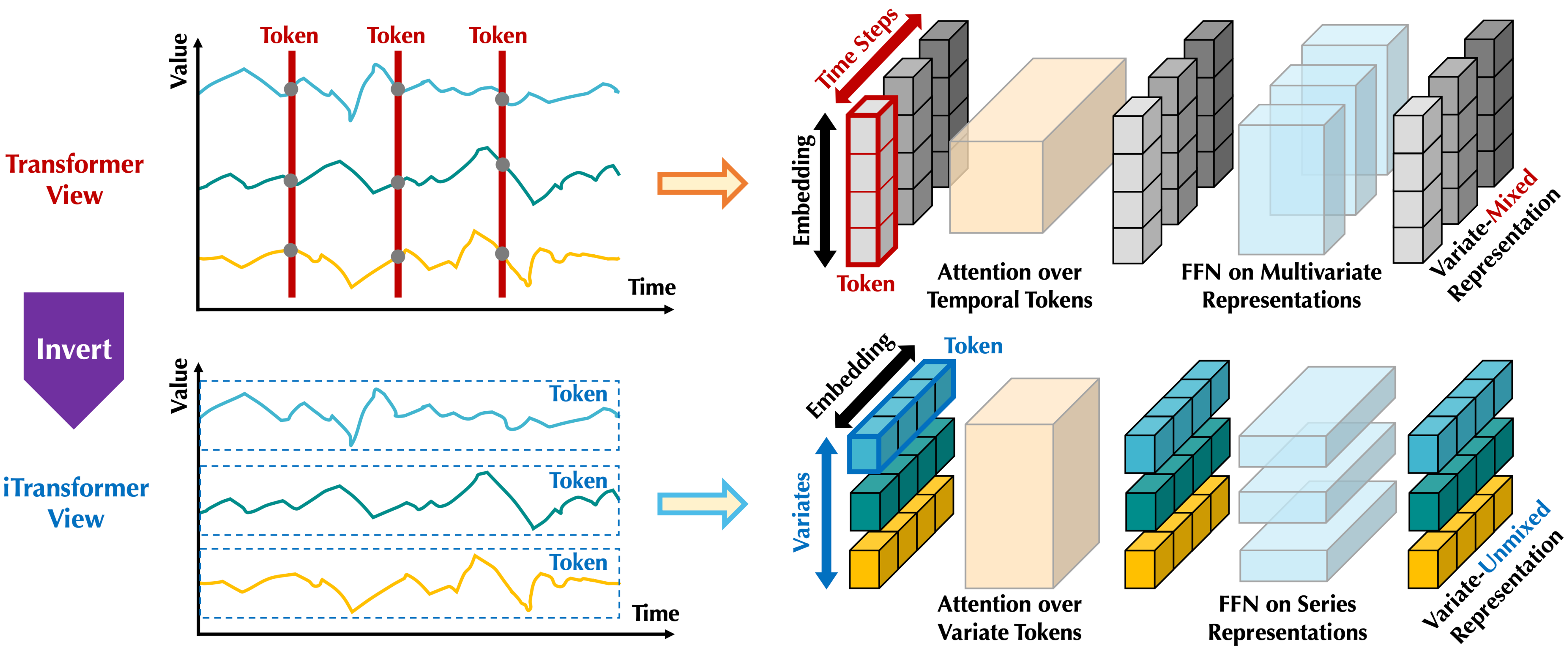
- Excessively local receptive field
- Inconsistent scale and distribution
- Variate-mixed representations
- Inherent lags between variates…





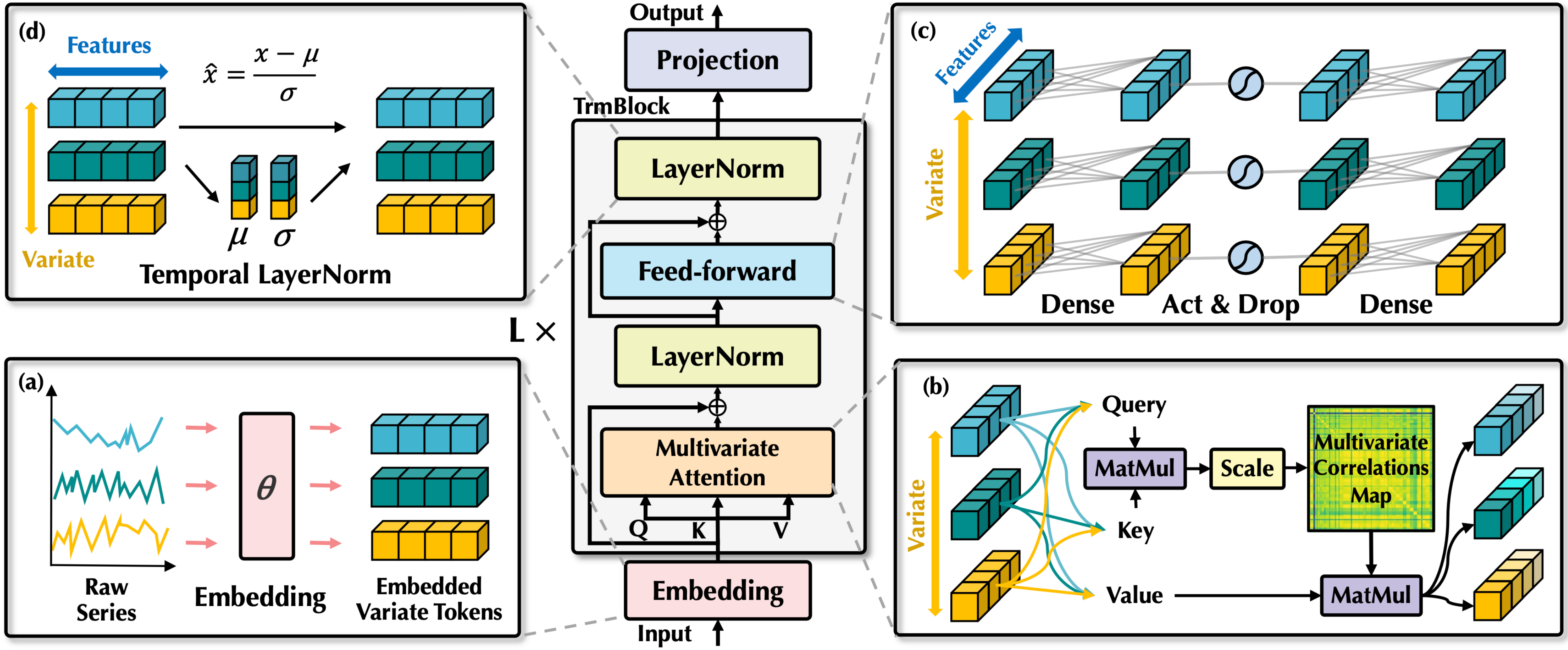


Motivation



iTransformer regards multivariate series **invertedly** without any modular modification

iTransformer

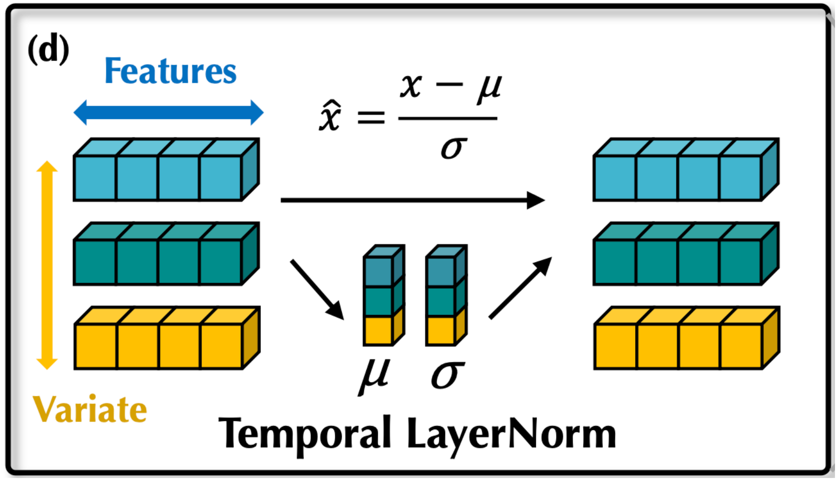


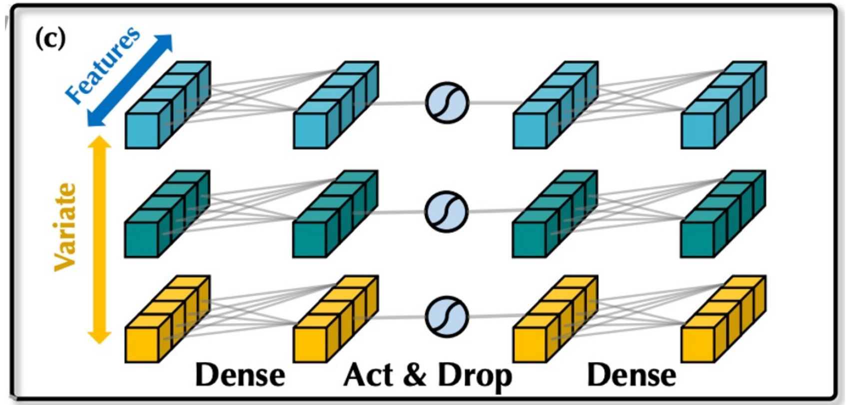
Encoder-only Arch.

$$\begin{aligned} \mathbf{h}_n^0 &= \text{Embedding}(\mathbf{X}_{:,n}), \\ \mathbf{H}^{l+1} &= \text{TrmBlock}(\mathbf{H}^l), \\ \hat{\mathbf{Y}}_{:,n} &= \text{Projection}(\mathbf{h}_n^L). \end{aligned}$$

- Time series of individual variate as the Variate Token
- LayerNorm and FFN for Variate-centric Representations
- Multivariate Correlations are captured by self-attention

Transformer Modules





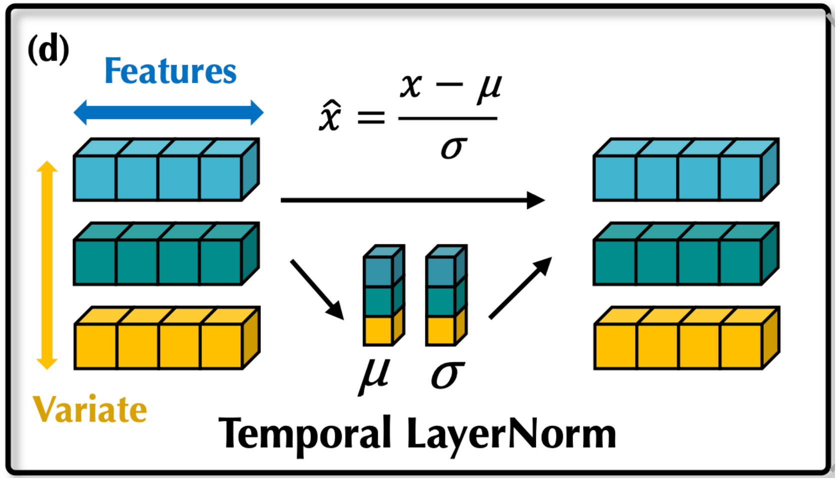
Layer normalization (within Variate Tokens)

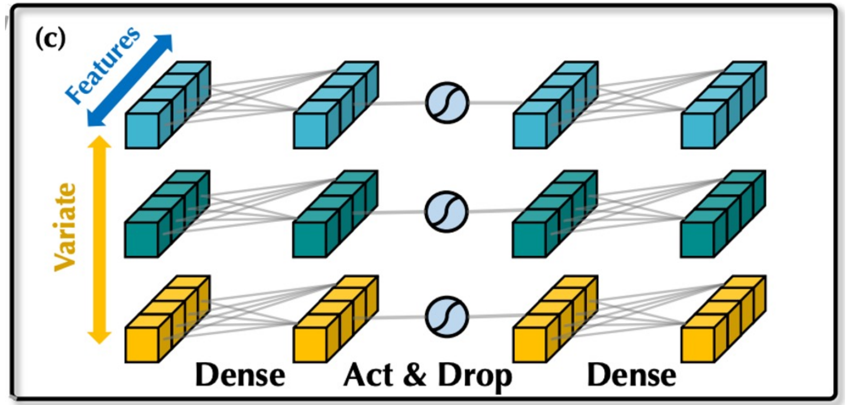
$$\text{LayerNorm}(\mathbf{H}) = \left\{ \frac{\mathbf{h}_n - \text{Mean}(\mathbf{h}_n)}{\sqrt{\text{Var}(\mathbf{h}_n)}} \,\middle|\, n = 1, \cdots, N \right\}$$

Mitigate variate discrepancies in scaling and distribution

Instead, time-unaligned events are merged and the obtained Temporal Tokens can be over-smoothed

Transformer Modules





Layer normalization (within Variate Tokens)

$$\mathrm{LayerNorm}(\mathbf{H}) = \left\{ \frac{\mathbf{h}_n - \mathrm{Mean}(\mathbf{h}_n)}{\sqrt{\mathrm{Var}(\mathbf{h}_n)}} \middle| n = 1, \cdots, N \right\}$$

Mitigate variate discrepancies in scaling and distribution

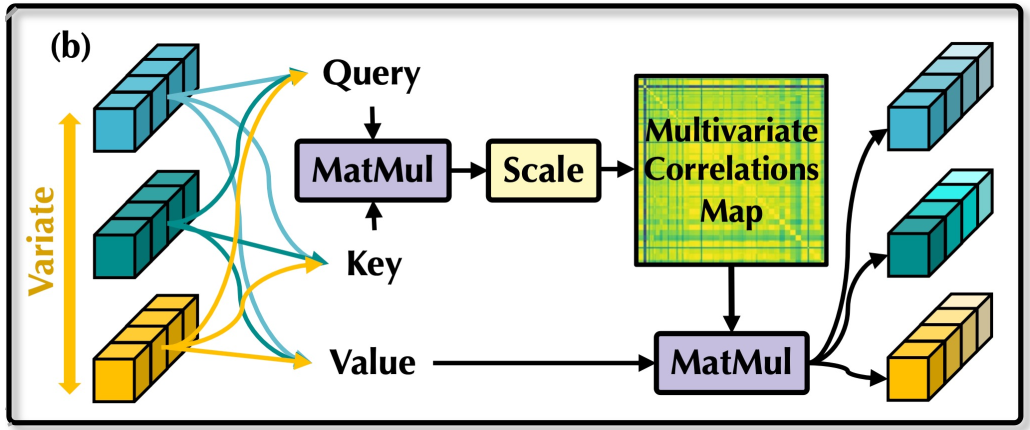
Instead, time-unaligned events are merged and the obtained Temporal Tokens can be over-smoothed

Feed-forward network (within Variate Tokens)

- Learns temporal representation
- Describe intrinsic properties of time series
- Transferable representation across variates

Naturally captured: Nonlinear temporal representation under Channel Independence

Module Reflections

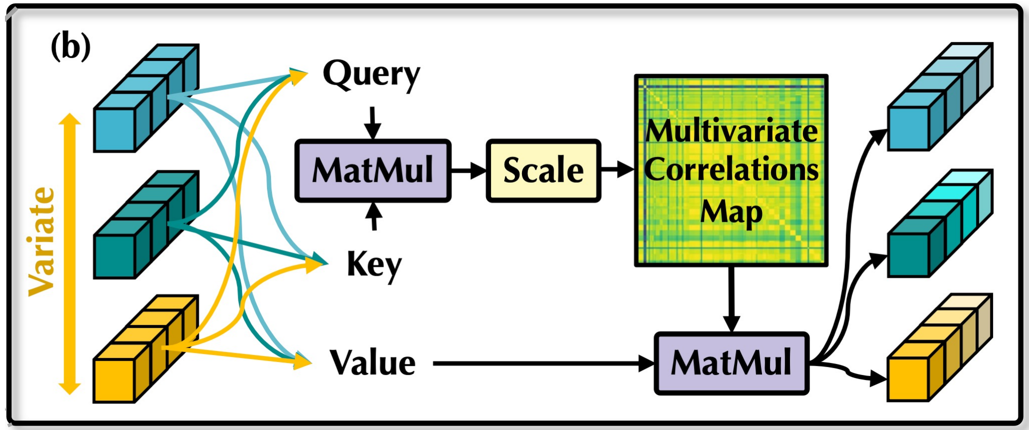


Self-attention (among Variate Tokens)

$\mathbf{H} = \{\mathbf{h}_0, \ldots, \mathbf{h}_N\}$ N - Variate number

$\mathbf{q}_i, \mathbf{k}_j$ - Query and key of Variate Tokens

Module Reflections



Self-attention (among Variate Tokens)

$\mathbf{H} = \{\mathbf{h}_0, \dots, \mathbf{h}_N\}$ N - Variate number

$\mathbf{q}_i, \mathbf{k}_j$ - Query and key of Variate Tokens

Pearson Correlation coefficients:

$$\rho_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Elements of Attention:

$$\mathbf{A}_{i,j} = (\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k})_{i,j} \propto \mathbf{q}_i^\top \mathbf{k}_j$$

$\mathbf{A} \in \mathbb{R}^{N \times N}$ - Multivariate Correlations

Highly correlated tokens will be more weighted with the Value $\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$

Time Series Forecasting

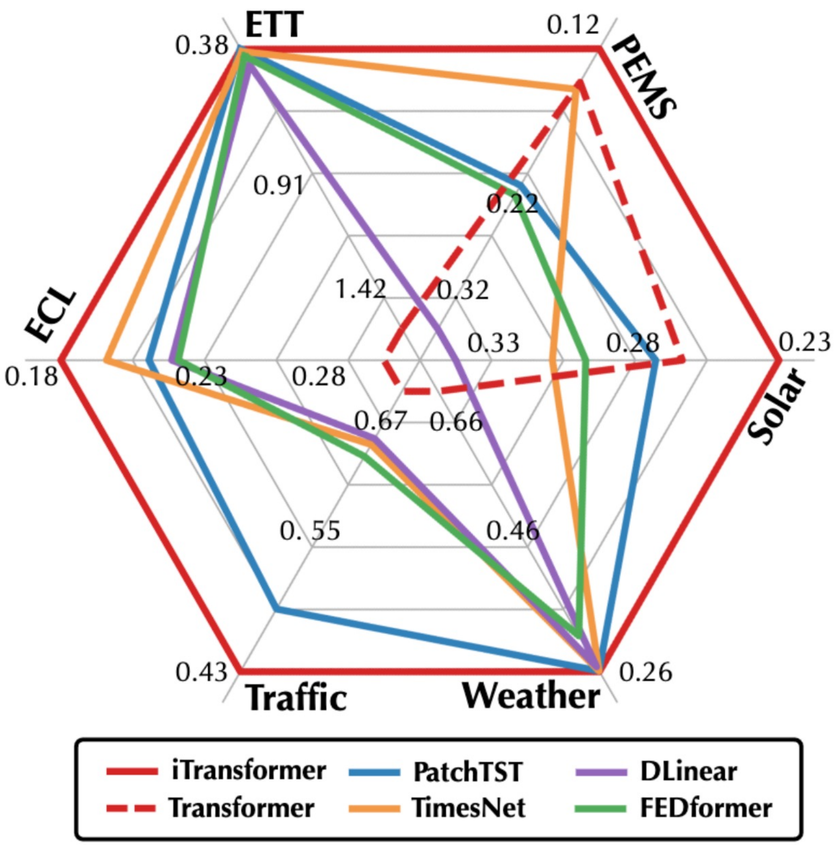
7 Benchmark (13 Datasets, 52 Prediction Settings)

Models	iTransformer (Ours)		RLinear (2023)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		TimesNet (2023)		DLinear (2023)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	0.178	0.270	0.219	0.298	0.205	0.290	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300
ETT (Avg)	0.383	0.399	0.380	0.392	0.381	0.397	0.685	0.578	0.482	0.470	0.391	0.404	0.442	0.444
Exchange	0.360	0.403	0.378	0.417	0.367	0.404	0.940	0.707	0.370	0.413	0.416	0.443	0.354	0.414
Traffic	0.428	0.282	0.626	0.378	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383
Weather	0.258	0.278	0.272	0.291	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317
Solar-Energy	0.233	0.262	0.369	0.356	0.270	0.307	0.641	0.639	0.347	0.417	0.301	0.319	0.330	0.401
PEMS (Avg)	0.119	0.218	0.514	0.482	0.217	0.305	0.220	0.304	0.375	0.440	0.148	0.246	0.320	0.394

Market Datasets (Server Load Prediction of Ant Group)

Models	iTransformer (Ours)		RLinear (2023)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		TimesNet (2023)		DLinear (2023)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Merchant	0.072	0.147	0.152	0.247	0.084	0.171	0.117	0.181	0.187	0.289	0.093	0.184	0.110	0.206
Wealth	0.345	0.289	0.585	0.461	0.394	0.326	0.429	0.288	0.595	0.481	0.360	0.318	0.501	0.412
Finance	0.184	0.216	0.395	0.336	0.231	0.248	5.333	0.618	0.987	0.442	0.516	0.308	0.765	0.372
Terminal	0.065	0.150	0.180	0.286	0.077	0.179	0.071	0.162	0.216	0.311	0.080	0.179	0.106	0.210
Payment	0.072	0.144	0.143	0.245	0.084	0.174	0.207	0.179	0.208	0.278	0.105	0.182	0.116	0.200
Customer	0.094	0.150	0.214	0.261	0.118	0.180	0.309	0.194	0.308	0.307	0.142	0.191	0.184	0.219

Averaged MSE (4 prediction lengths)



Achieve state-of-the-art on MTSF

Excel at high-dimensional series: ECL, Traffic, Solar...

Framework Generality

Models		Transformer (2017)		Reformer (2020)		Informer (2021)		Flowformer (2022)		Flashformer (2022)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	Original	0.277	0.372	0.338	0.422	0.311	0.397	0.267	0.359	0.285	0.377
	+Inverted	**0.178**	**0.270**	**0.208**	**0.301**	**0.216**	**0.311**	**0.210**	**0.293**	**0.206**	**0.291**
	Promotion	35.6%	27.4%	38.4%	28.7%	30.5%	21.6%	21.3%	18.6%	27.8%	22.9%
Traffic	Original	0.665	0.363	0.741	0.422	0.764	0.416	0.750	0.421	0.658	0.356
	+Inverted	**0.428**	**0.282**	**0.647**	**0.370**	**0.662**	**0.380**	**0.524**	**0.355**	**0.492**	**0.333**
	Promotion	35.6%	22.3%	12.7%	12.3%	13.3%	8.6%	30.1%	15.6%	25.2%	6.4%
Weather	Original	0.657	0.572	0.803	0.656	0.634	0.548	0.286	0.308	0.659	0.574
	+Inverted	**0.258**	**0.279**	**0.248**	**0.292**	**0.271**	**0.330**	**0.266**	**0.285**	**0.262**	**0.282**
	Promotion	60.2%	50.8%	69.2%	55.5%	57.3%	39.8%	7.2%	7.7%	60.2%	50.8%

- Inverting can consistently improve various Transformers
- Take advantage of booming efficient attention mechanisms

Prediction Accuracy

Transformer
↑ 38.9%

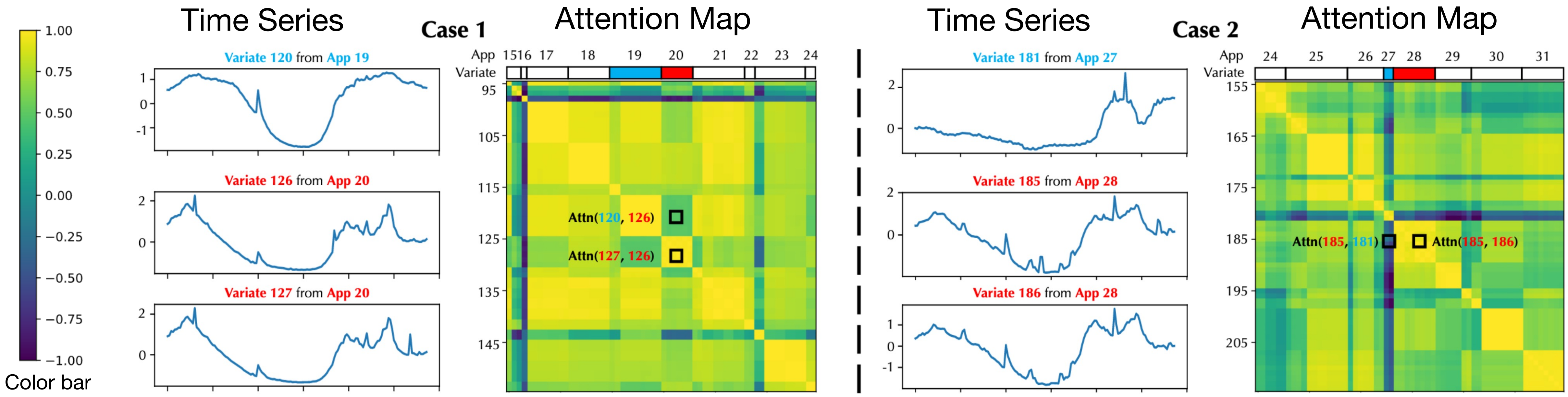
Reformer
↑ 36.1%

Informer
↑ 28.5%

Flowformer
↑ 16.8%

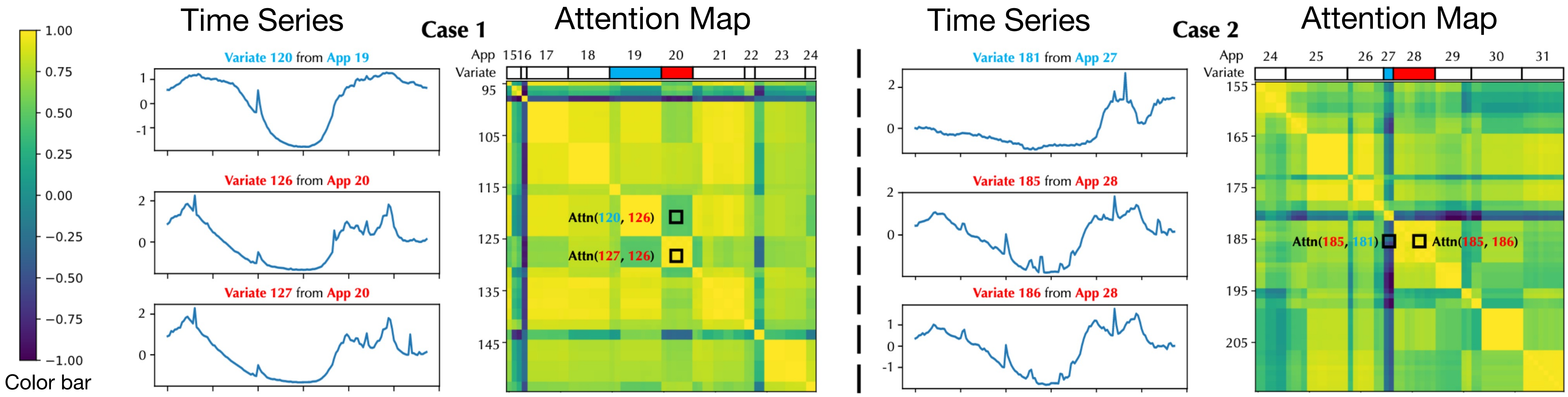
Flashformer
↑ 32.2%

Multivariate Correlations



Market Dataset: each variate represents the monitored series of a service interface of a kind

Multivariate Correlations

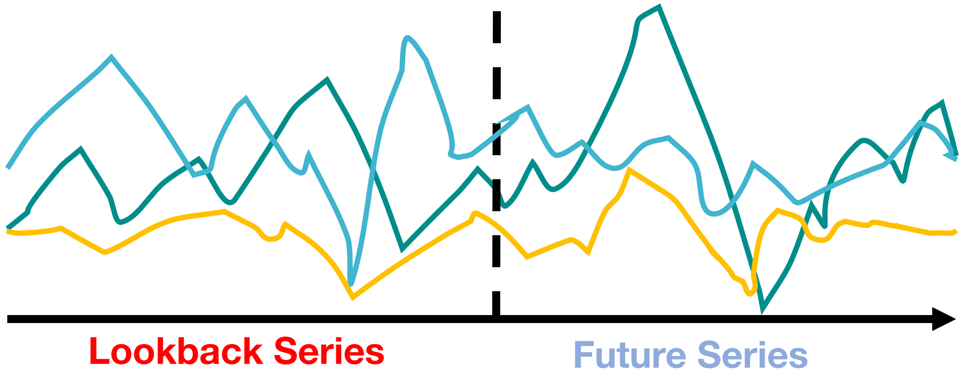


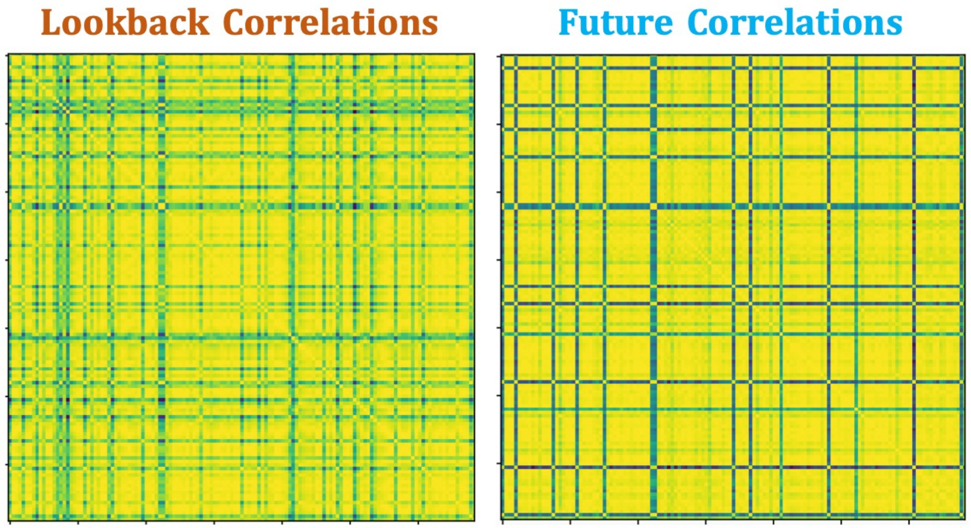
Market Dataset: each variate represents the monitored series of a service interface of a kind

- Partitions in the learned attention map, indicating the grouping of variates
- The learned attention map reveals the correlations between the variates

Multivariate Correlations

Solar-Energy Dataset: distinct multivariate correlations in the lookback and future series





Calculated from raw series

$$\rho_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

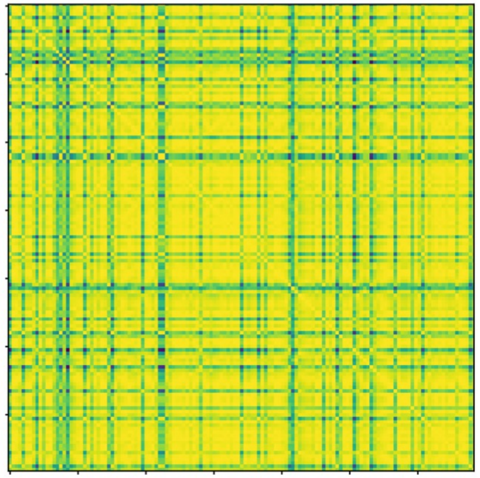
Multivariate Correlations

Solar-Energy Dataset: distinct variate correlations in the lookback and future series

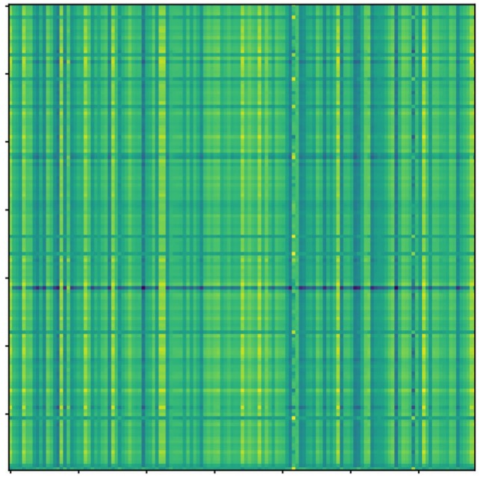
Attention map can reflect the correlation between the variates

In the shallow layer, the map share similarities to the correlations of lookback series

Lookback Correlations



$$\rho_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$



$$\mathbf{A}_{i,j} = (\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k})_{i,j} \propto \mathbf{q}_i^\top \mathbf{k}_j$$

Score Map of Layer 1

Learned by iTransformer

Multivariate Correlations

Solar-Energy Dataset: distinct variate correlations in the lookback and future series

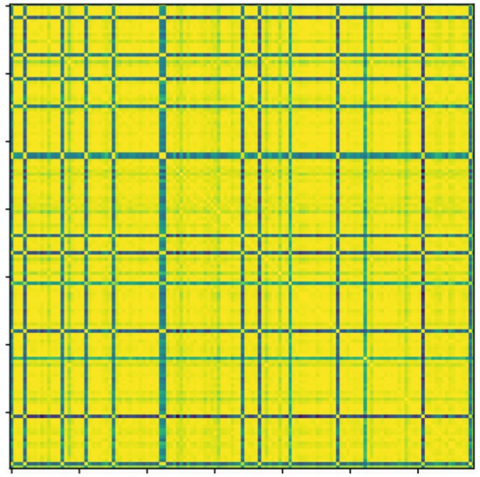
$$\rho_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

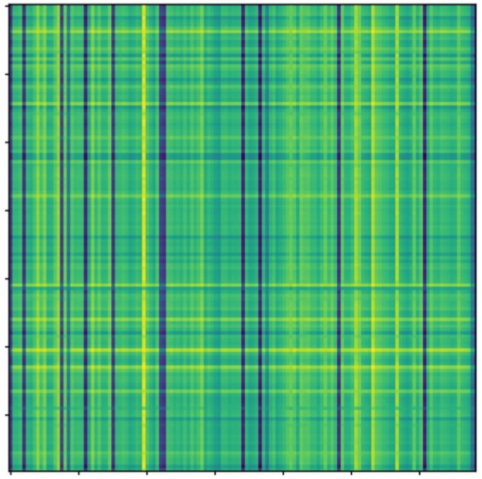
Attention map reflect the correlation between the variates

In the deep layer, the map share similarities to the correlations of future series

$$\mathbf{A}_{i,j} = (\mathbf{Q}\mathbf{K}^\top/\sqrt{d_k})_{i,j} \propto \mathbf{q}_i^\top \mathbf{k}_j$$

Future Correlations





Score Map of Layer L

Learned by iTransformer

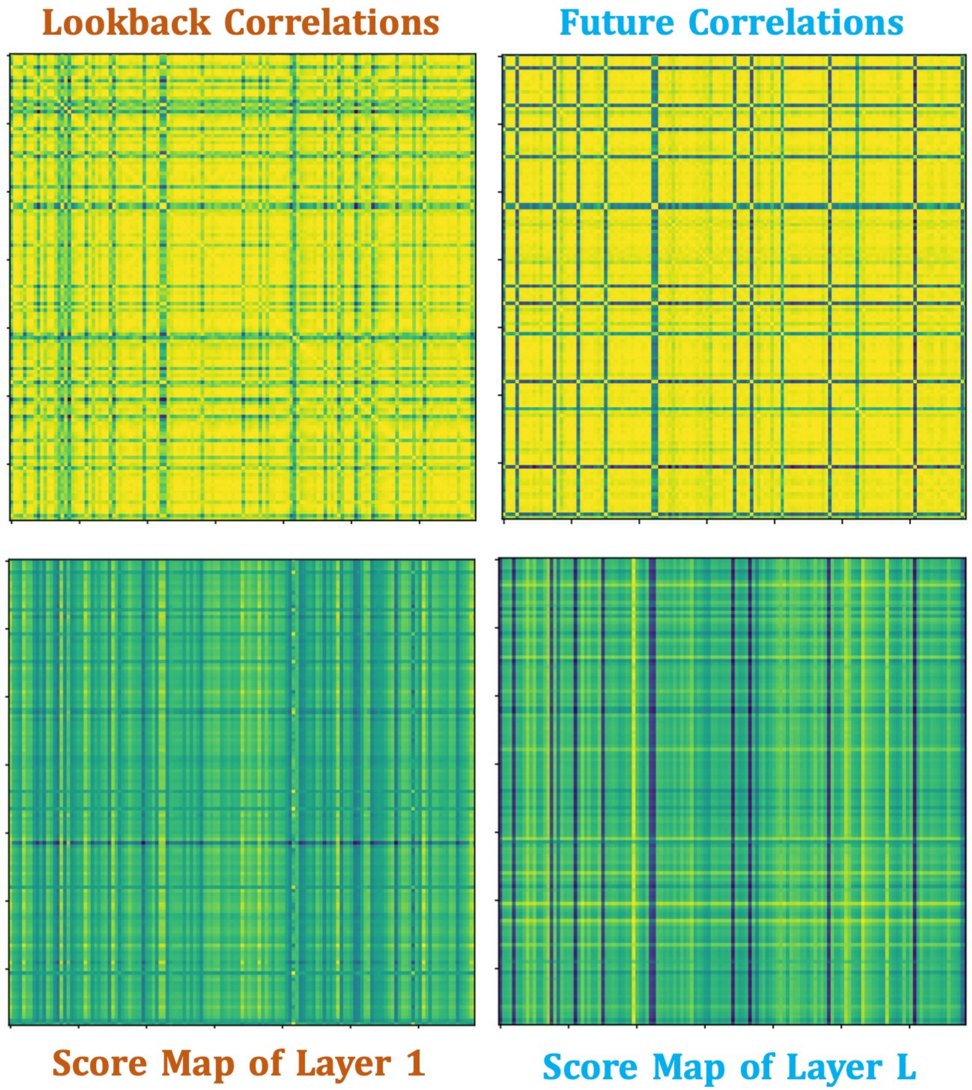
Multivariate Correlations

Solar-Energy Dataset: distinct variate correlations in the lookback and future series

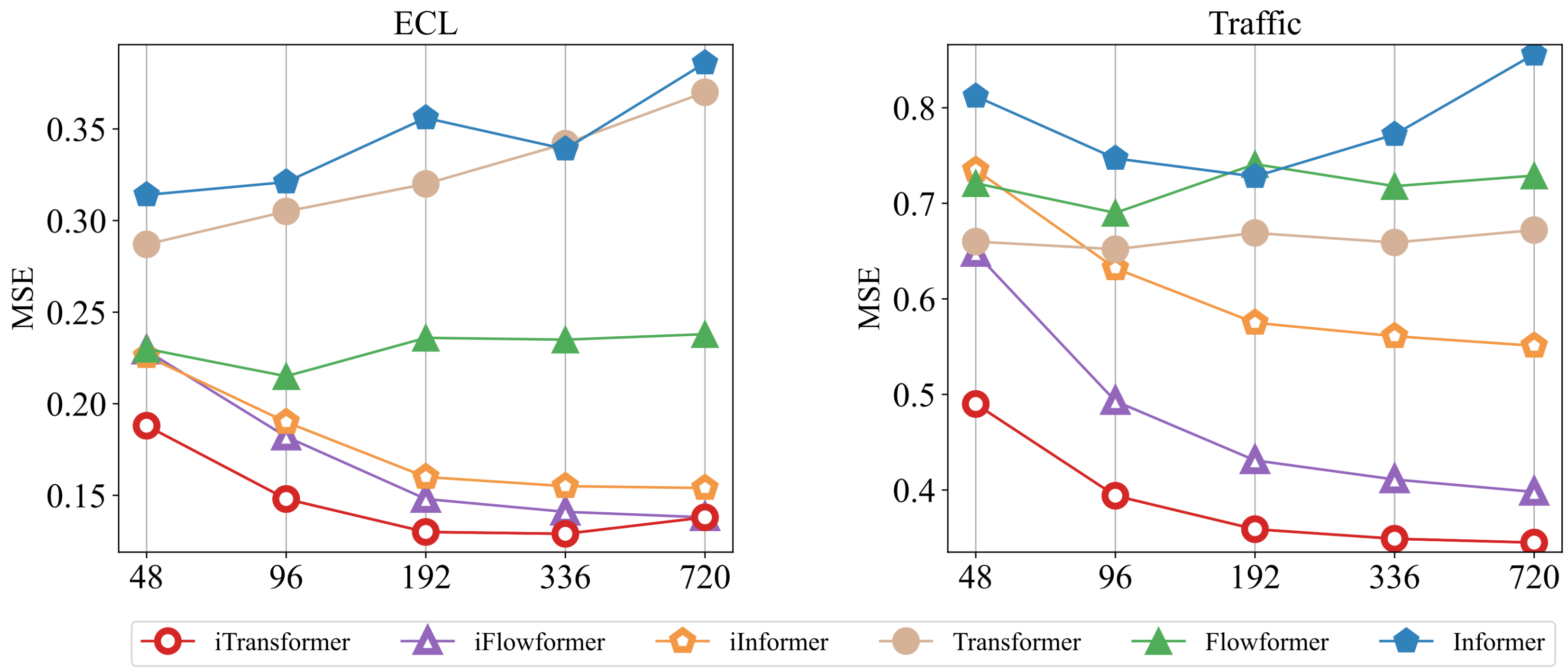
Attention scores reflect the correlation between the variates

Inverting empowers

- Attention: Interpretable variate correlating
- FFN & LN: Encoding Variate Tokens and decoding them for the prediction



Prolonged Lookback Length

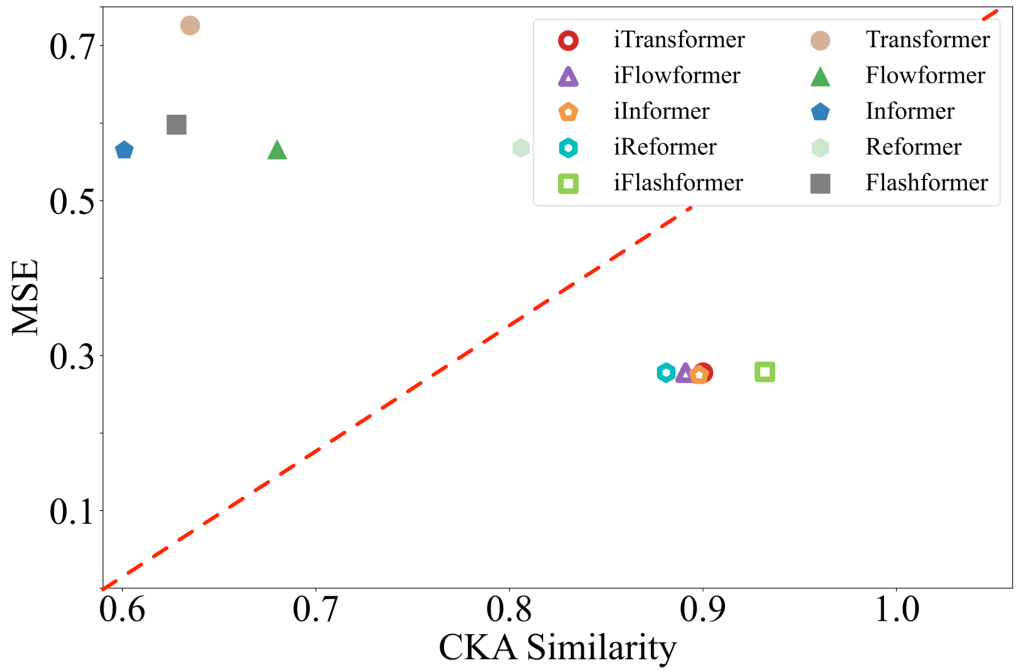


Previous work found that Transformer-based forecasters does not necessarily improve with enlarged lookback wnidow

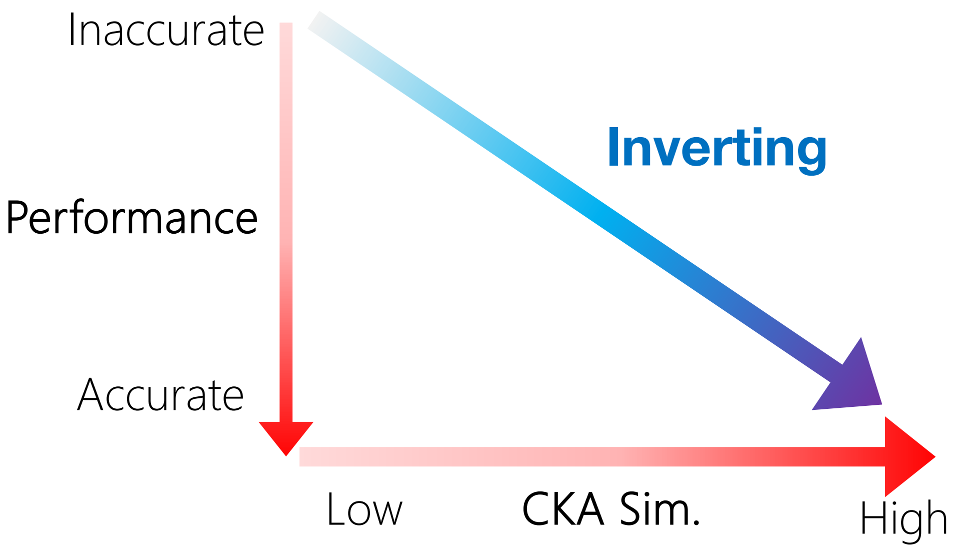
Performance of iTransformer is generally improving with more lookback observations

Nie et al. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *ICLR, 2023.*

Representation Analysis



Forecasting Performance V.S Centered Kernel Alignment



Previous work demonstrated that time series forecasting, as a low-level generative task, prefers the higher CKA similarity

Inverting learns favored series representations and thus achieves more accurate predictions.

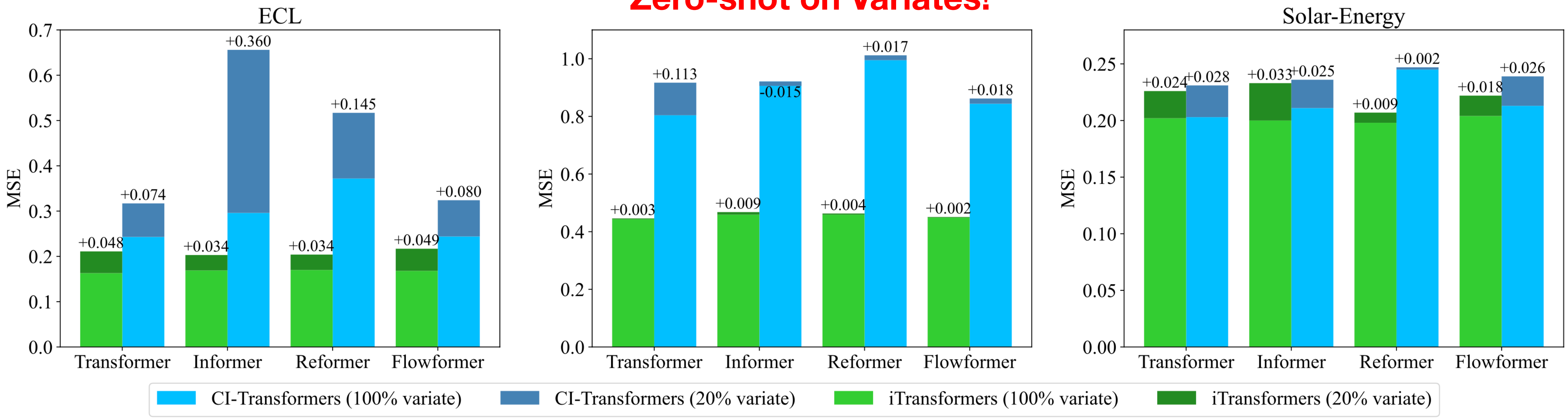
TimesNet: Temporal 2d-variation Modeling for General Time Series Analysis. *ICLR, 2023.*

Variate Generalization

Based on the **independent embedding** of Variate Tokens

iTransformer can forecast with arbitrary numbers of variates during inference

Zero-shot on Variates!



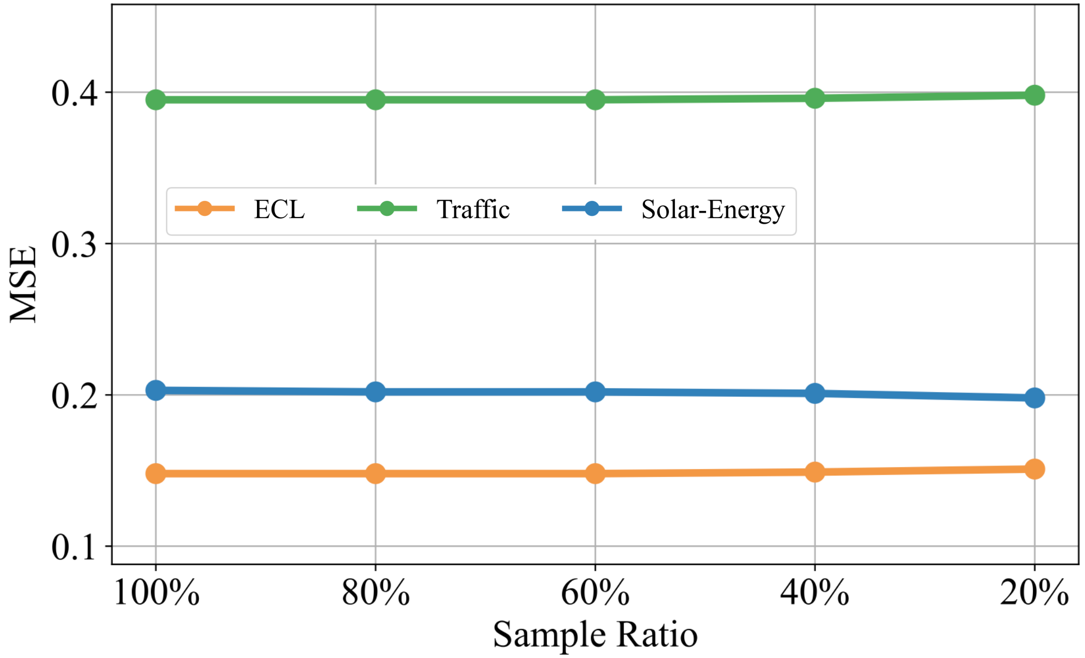
Similar to Channel Independence

iTransformers can be trained on partial variables and generalize well on unseen variates

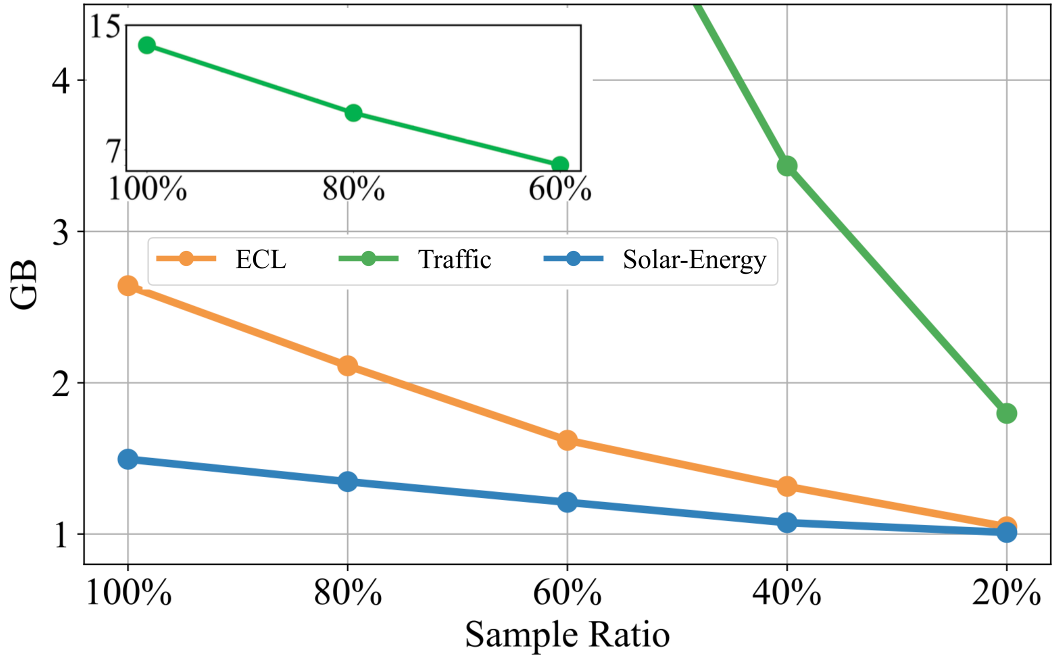
Nie et al. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *ICLR, 2023.*

Efficient Training

Performance



Memory Footprint

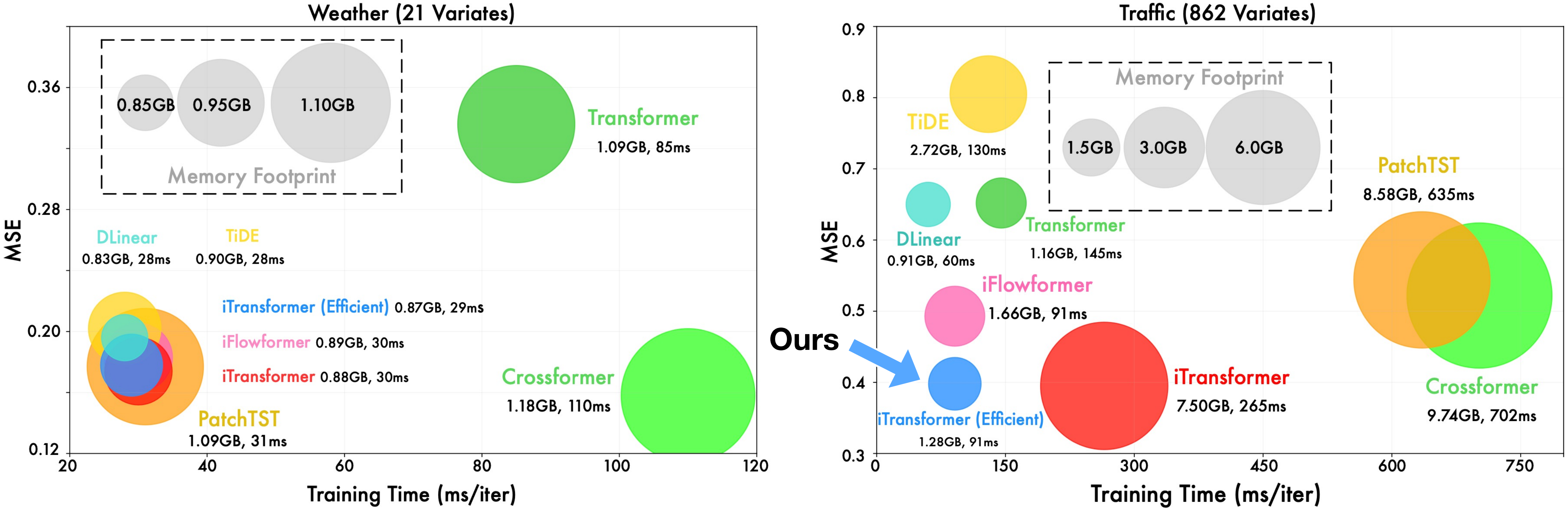


Based on the Variate Generalization Capability

We proposed an Efficient Training Strategy that trains sampled variates in each batch

Performance remains stable while memory footprints can be cut off significantly

Efficiency



- iTransformer exceeds other Transformers in datasets with a small number of variates
- Via Efficient Training, iTransformer shows strength in both Performance/Efficiency

Thank You!

liuyong21@mails.tsinghua.edu.cn

Code and datasets are available at https://github.com/thuml/iTransformer