Unified Transferability Metrics for Time Series Foundation Models

Weiyang Zhang, Xinyang Chen✉, Xiucheng Li, Kehai Chen, Weili Guan, Liqiang Nie
Harbin Institute of Technology (Shenzhen)

Summary

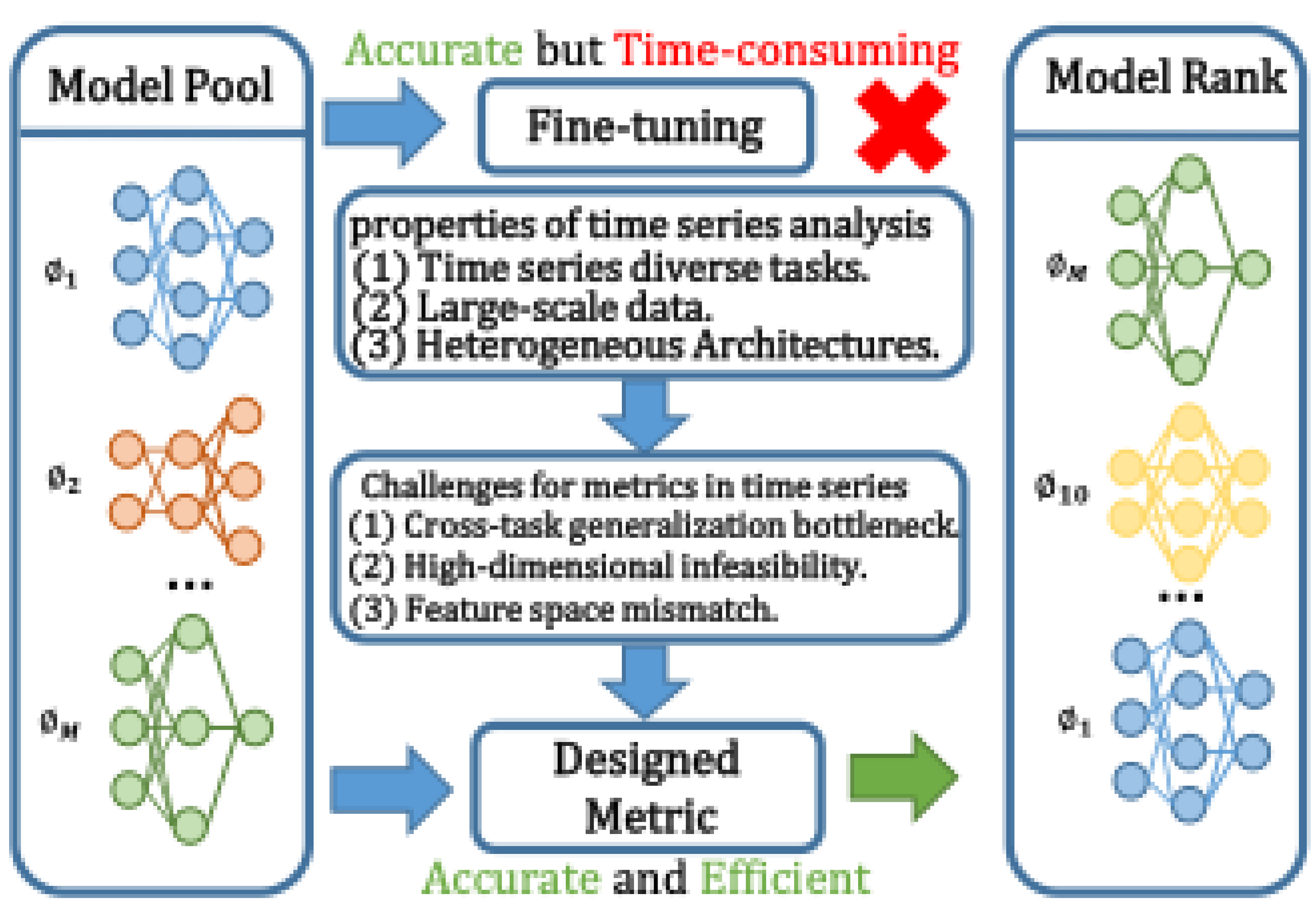
- Empirical Insight: A good time series pre-trained model can effectively capture long-term dependencies and key temporal patterns, while different downstream tasks have distinct feature requirements.
- TEMPLATE:A flexible and generalizable method, supporting both classification and regression tasks.
- General Evaluations : TEMPALTE achieve state-of-the-art performance on all downstream tasks, including classification, forecasting, imputation, and anomaly detection !

Background

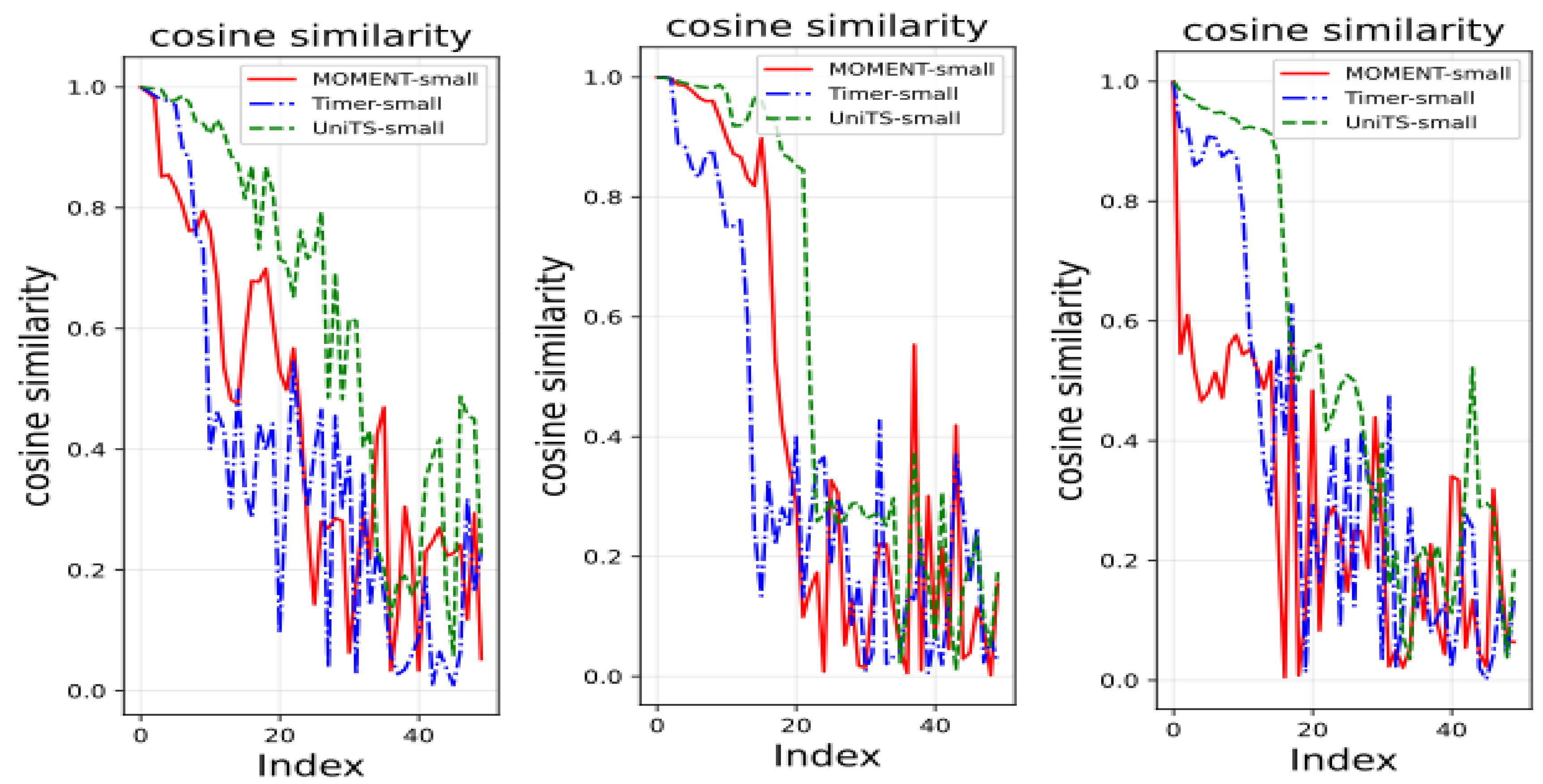
- Pre-trained models in the field of time series are constantly increasing, and a large number of pre-trained models are already available on open-source platforms.
- No single pre-trained model can perform well on all time series downstream tasks, thus how to quickly select the pre-trained model suitable for downstream tasks without fine-tuning has become an urgent problem to be solved.

Challenges

- Cross-task generalization bottleneck: Mainstream time series tasks exhibit diversity, and the designed metrics need to achieve cross-task adaptability.
- High-dimensional infeasibility: Downstream datasets feature large-scale characteristics. Designed metrics must balance efficiency and accuracy.
- Feature space mismatch: Heterogeneity of model architectures and differences approaches to handling inter-channel dependencies result in discrepancies.



Motivation



- By comparing the changes in the feature matrix before and after fine-tuning, it is found that larger singular values are more stable than smaller ones.

Methods

- **Preliminary**

The feature extracted by the l-th layer of the pre-trained model $\Phi_m(\cdot)$ is denoted as $\mathbf{H}^l$, where $\mathbf{H}^l = \Phi_m(\mathbf{X}) \in \mathbb{R}^{N\times d}$ and d is the feature dimension.

- **Dependency Learning Score**
  - Obtain the feature matrix of the trend component via trend decomposition.

$$\mathbf{T} = \phi_m(trend(\mathbf{X})) \quad (1)$$

  - Perform SVD to decompose the features.

$$\mathbf{H} = U_h \Sigma_h V_h^T, \mathbf{T} = U_t \Sigma_t V_t^T \quad (2)$$

  - Quantify the model's capability in capturing long-term dependencies.

$$S_{dl} = \frac{Conv(u_h, u_t)}{\lambda_h \lambda_t} \quad (3)$$

- **Pattern Learning Score** :
  - Quantify the model's capability in learn primary temporal patterns.

$$S_{pl} = \frac{\sigma_t}{||\mathbf{T}||_*} \quad (4)$$

- **Task Adaptation Score** :
  - Quantify the model's capability in adapting to downstream tasks .

$$S_{ta} = \frac{HSIC(K, L)}{HSIC(K, K)HSIC(L, L)}. \quad (5)$$

Experiment Results

Table: Classification Benchmark Performance (Weighted Kendall's τw) of different methods

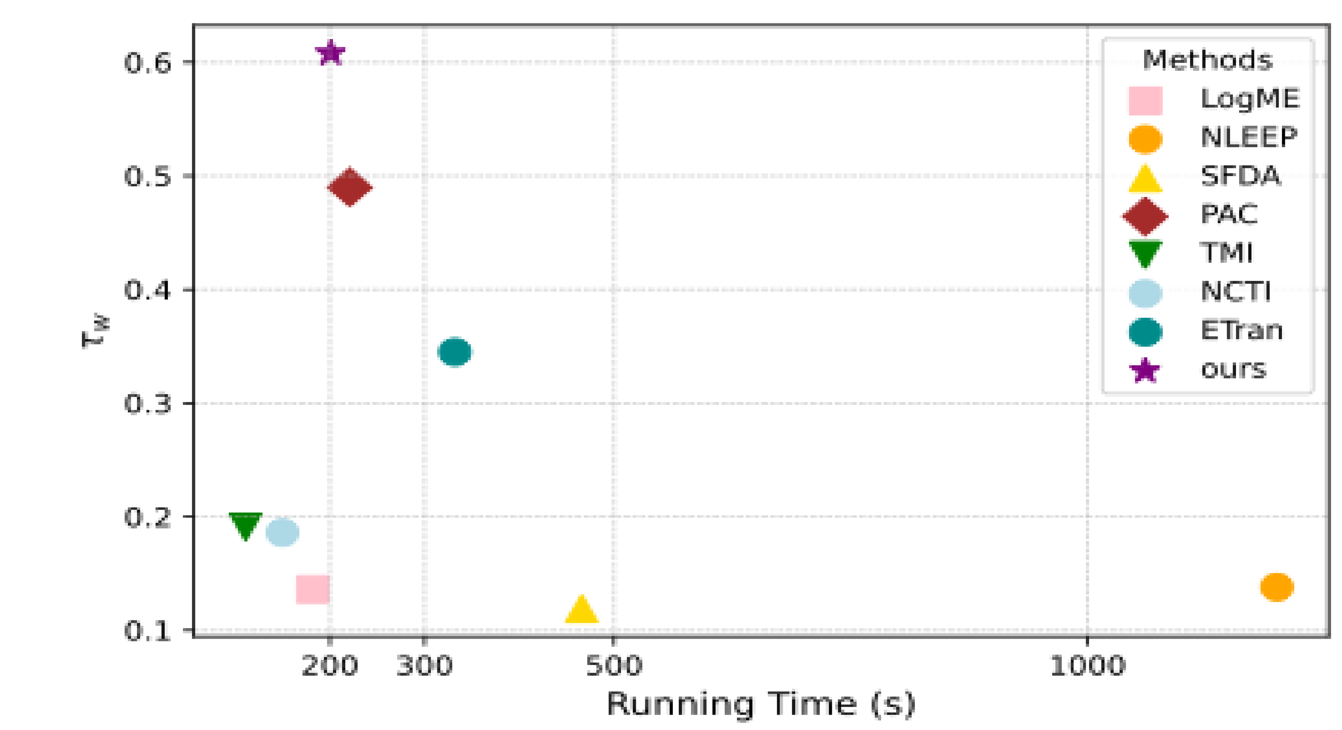
Datasets	LogME	NLEEP	SFDA	PACTran	TMI	NCTI	ETran	RetMMD	TEMPLATE
EthanolConcentration	0.567	0.432	0.120	0.488	-0.430	-0.32	0.686	0.512	**0.724**
FaceDetection	-0.203	0.092	**0.598**	0.306	-0.600	0.109	-0.359	0.310	0.597
Handwriting	-0.445	-0.104	0.314	0.596	0.768	0.700	0.478	0.365	**0.822**
JapaneseVowels	0.231	0.213	0.021	0.306	0.302	0.340	-0.196	**0.654**	0.447
PEMS-SF	-0.472	-0.612	-0.312	0.306	-0.300	0.053	0.076	**0.520**	0.470
SelfRegulationSCP1	0.356	0.529	0.459	0.619	0.300	0.310	**0.651**	0.450	0.484
SelfRegulationSCP2	0.268	0.241	-0.198	0.457	0.455	0.450	**0.667**	0.397	0.551
SpokenArabicDigits	0.342	0.321	-0.367	**0.744**	0.647	-0.210	0.479	0.201	0.637
UWaveGestureLibrary	0.584	0.127	0.440	0.592	0.576	0.245	0.624	0.362	**0.719**
Average	0.136	0.138	0.119	0.490	0.191	0.186	0.345	0.419	**0.608**

Table: Forecasting Benchmark Performance (Weighted Kendall's τw) of different methods

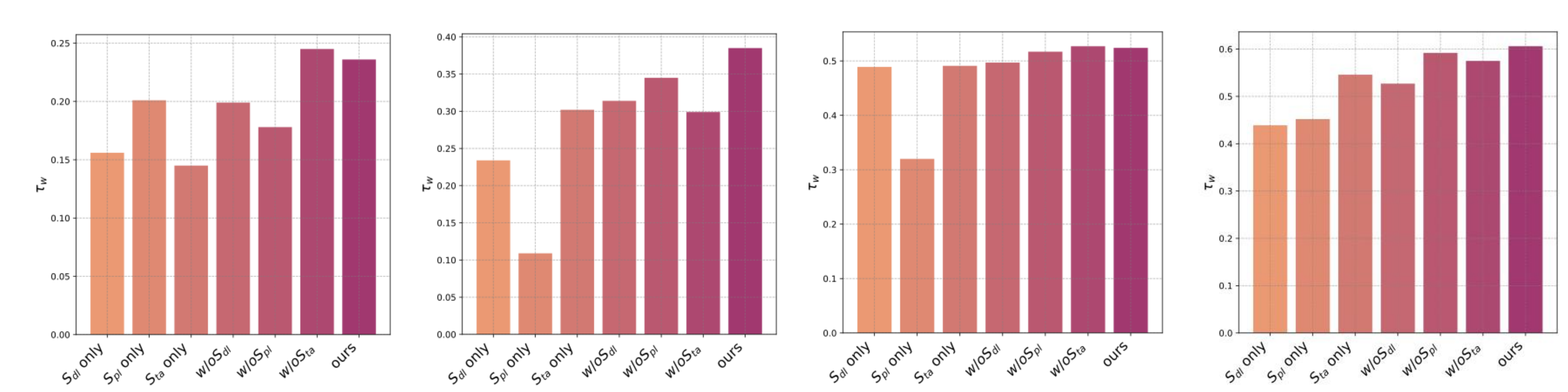
Methods	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Electricity	Traffic	Average
LogME	0.215	0.167	0.400	0.565	0.114	-0.130	-	0.190
ETran	0.138	**0.212**	0.189	0.351	**0.474**	0.302	0.192	0.265
Ours	**0.240**	-0.003	**0.518**	**0.576**	0.412	**0.361**	**0.432**	**0.362**

Analysis

- Time complexity analysis: TEMPLATE strikes a high degree of balance between efficiency and accuracy.



- Ablation Study: All three metrics achieve positive ranking correlations, and their combination yields the highest average ranking correlation.



School of Computer Science and Technology - Harbin Institute of Technology (Shenzhen) – China
Mail: zhangweiyang12138@gmail.com