TS2Vec: Towards Universal Representation of Time Series

Machine Learning in Practice Reading Group

Duke B&B

February 2, 2024

Presented by Yuankang Zhao

Section 1: Introduction

Limitations of current method

- Instance-level representations may not be suitable for tasks that need fine-grained representations.
- Eg. time series forecasting and anomaly detection
- Current methods fail to featurizes time series at different scales to capture scale-invariant information
- Multiscale features(daily, monthly) may provide different levels of semantics and improve the generalization capability of learned representations.
- Current method inspired by experiences in CV and NLP domains strong inductive bias such as transformation-invariance and cropping-invariance

Purpose: Purpose a universal framework for learning representations of time series in all semantic levels.

Problem definition

- Given a set of time series $X = \{x_1, x_2, \ldots, x_N\}$ of N instances
- The goal is to learn a nonlinear embedding function f_θ that maps each x_i to its representation r_i that best describes itself.
- The input time series x_i has dimension $T \times F$, where T is the sequence length and F is the feature dimension.
- The representation $r_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,T}\}$ contains representation vectors $r_{i,t} \in \mathbb{R}^K$ for each timestamp t, where K is the dimension of representation vectors.

Model Architecture

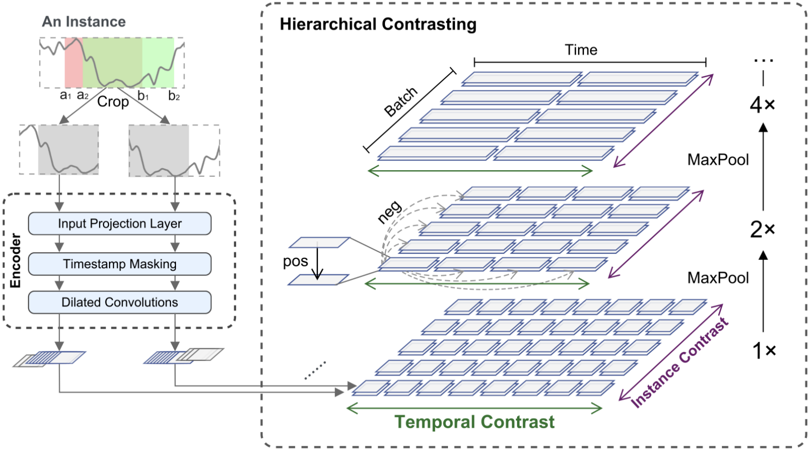
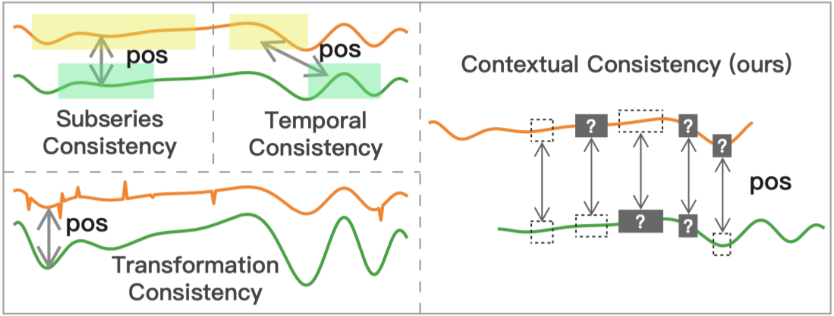


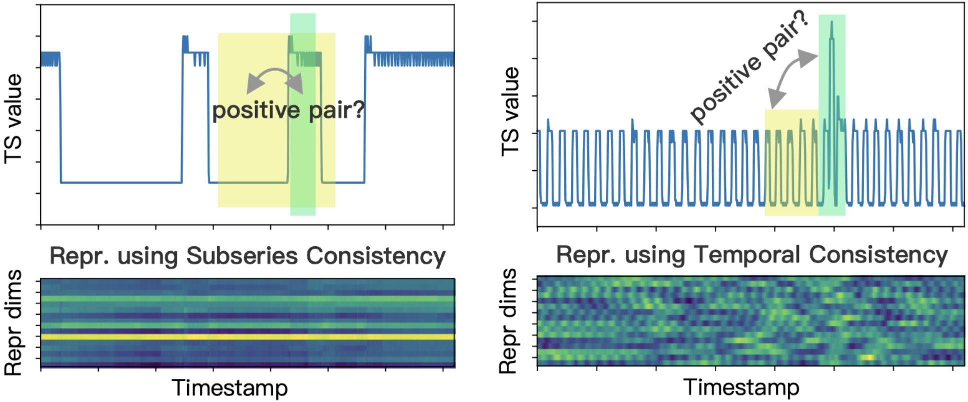
Figure 1: The proposed architecture of TS2Vec. Although this figure shows a univariate time series as the input example, the framework supports multivariate input. Each parallelogram denotes the representation vector on a timestamp of an instance.

Previous strategies of constructing positive pairs

- Subseries consistency: encourages the representation of a time series to be closer to its sampled subseries.
- Temporal consistency: enforces the local smoothness of representations by choosing adjacent segments as positive samples.
- Transformation consistency: augments input series by different transformations, such as scaling, permutation, etc., encouraging the model to learn transformation-invariant representations.



Previous strategies may fail in some cases



(a) Level shifts. (b) Anomalies.

Section 3: Methods

Purposed strategy of generating positive samples: Contextual Consistency

Timestamp Masking

- randomly mask the timestamps of an instance to produce a new context view
- masks the latent vector $z_i = \{z_{i,t}\}$ after the Input Projection Layer along the time axis with a binary mask m
- $m \sim Bernoulli(0.5)$

Random Cropping

- For any time series input $x_i \in \mathbb{R}^{T \times F}$
- TS2Vec randomly samples two overlapping time segments $[a_1, b_1], [a_2, b_2]$ such that $0 < a_1 \leq a_2 \leq b_1 \leq b_2 \leq T$.
- The contextual representations on the overlapped segment $[a_2, b_1]$ should be consistent for two context views.
- random cropping helps learn position-agnostic representations and avoids representation collapse.

Timestamp masking and random cropping are only applied in the training phase.

Hierarchical Contrasting

Temporal Contrastive Loss: This loss function takes representations from the same timestamp as positives and from different timestamps as negatives to learn discriminative features over time.

$$\ell_{i,t}^{\text{temp}} = -\log \frac{\exp(r_{i,t} \cdot r'_{i,t})}{\sum_{t' \in \Omega} \exp(r_{i,t} \cdot r'_{i,t'}) + \mathbb{I}_{[t \neq t']} \exp(r_{i,t} \cdot r'_{i,t'})}$$

Instance-wise Contrastive Loss: This loss uses representations of other time series at the same timestamp in the same batch as negative samples.

$$\ell_{i,t}^{\text{inst}} = -\log \frac{\exp(r_{i,t} \cdot r'_{i,t})}{\sum_{j=1}^{B} \exp(r_{i,t} \cdot r'_{j,t}) + \mathbb{I}_{[i \neq j]} \exp(r_{i,t} \cdot r'_{j,t})}$$

Overall Loss: The final loss is a combination of the temporal and instance-wise losses, averaged over all time series and timestamps.

$$L_{\text{dual}} = \frac{1}{NT} \sum_{i} \sum_{t} (\ell_{i,t}^{\text{temp}} + \ell_{i,t}^{\text{inst}})$$

These contrastive losses complement each other, capturing both user-specific characteristics and dynamic trends over time.

Section 3: Methods

Hierarchical Contrasting

Algorithm 1: Calculating the hierarchical contrastive loss

procedure HIERLOSS(r, r')
    $\mathcal{L}_{hier} \leftarrow \mathcal{L}_{dual}(r, r')$;
    $d \leftarrow 1$;
    while $\text{time\_length}(r) > 1$ **do**
        // The maxpool1d operates along the time axis.
        $r \leftarrow \text{maxpool1d}(r, \text{kernel\_size} = 2)$;
        $r' \leftarrow \text{maxpool1d}(r', \text{kernel\_size} = 2)$;
        $\mathcal{L}_{hier} \leftarrow \mathcal{L}_{hier} + \mathcal{L}_{dual}(r, r')$;
        $d \leftarrow d + 1$;
    end while
    $\mathcal{L}_{hier} \leftarrow \mathcal{L}_{hier} / d$;
    return $\mathcal{L}_{hier}$
end procedure

Time Series Classification

- Classes are labeled on the entire time series (instance)
- The task requires the instance level representations. (Maxpooling overall timestamps)
- SVM classifier with RBF kernel is trained on top of the instance-level representations to make predictions.

	125 UCR datasets			29 UEA datasets		
Method	Avg. Acc.	Avg. Rank	Training Time (hours)	Avg. Acc.	Avg. Rank	Training Time (hours)
DTW	0.727	4.33	–	0.650	3.74	–
TNC	0.761	3.52	228.4	0.677	3.84	91.2
TST	0.641	5.23	17.1	0.635	4.36	28.6
TS-TCC	0.757	3.38	1.1	0.682	3.53	3.6
T-Loss	0.806	2.73	38.0	0.675	3.12	15.1
TS2Vec	**0.830 (+2.4%)**	**1.82**	**0.9**	**0.712 (+3.0%)**	**2.40**	**0.6**

Table 1: Time series classification results compared to other time series representation methods. The representation dimensions of TS2Vec, T-Loss, TS-TCC, TST and TNC are all set to 320 and under SVM evaluation protocol for fair comparison.

Time Series Classification

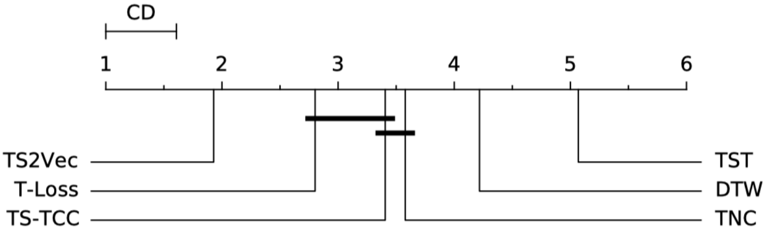


Figure 4: Critical Difference (CD) diagram of representation learning methods on time series classification tasks with a confidence level of 95%.

Time Series Forecasting

Problem Definition

- Given the last T_l observations $x_{t-T_l+1}, \ldots, x_t$
- forecasting task aims to predict the future H observations$(x_{t+1}, \ldots, x_{t+H}) = \hat{x}$
- A linear regression model with L_2 norm penalty that takes r_t, the representation of the last timestamp, as input to directly predict future values $\hat{x}$ to predict future observations.

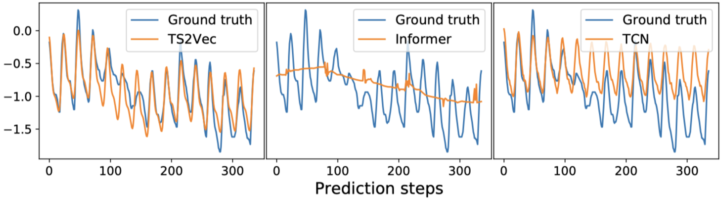


Figure 5: A prediction slice (H=336) of TS2Vec, Informer and TCN on the test set of $ETTh_2$.

Time Series Forecasting

Dataset	H	TS2Vec	Informer	LogTrans	N-BEATS	TCN	LSTnet
$ETTh_1$	24	**0.039**	0.098	0.103	0.094	0.075	0.108
	48	**0.062**	0.158	0.167	0.210	0.227	0.175
	168	**0.134**	0.183	0.207	0.232	0.316	0.396
	336	**0.154**	0.222	0.230	0.232	0.306	0.468
	720	**0.163**	0.269	0.273	0.322	0.390	0.659
$ETTh_2$	24	**0.090**	0.093	0.102	0.198	0.103	3.554
	48	**0.124**	0.155	0.169	0.234	0.142	3.190
	168	**0.208**	0.232	0.246	0.331	0.227	2.800
	336	**0.213**	0.263	0.267	0.431	0.296	2.753
	720	**0.214**	0.277	0.303	0.437	0.325	2.878
$ETTm_1$	24	**0.015**	0.030	0.065	0.054	0.041	0.090
	48	**0.027**	0.069	0.078	0.190	0.101	0.179
	96	**0.044**	0.194	0.199	0.183	0.142	0.272
	288	**0.103**	0.401	0.411	0.186	0.318	0.462
	672	**0.156**	0.512	0.598	0.197	0.397	0.639
Electric.	24	0.260	**0.251**	0.528	0.427	0.263	0.281
	48	**0.319**	0.346	0.409	0.551	0.373	0.381
	168	**0.427**	0.544	0.959	0.893	0.609	0.599
	336	**0.565**	0.713	1.079	1.035	0.855	0.823
	720	**0.861**	1.182	1.001	1.548	1.263	1.278
Avg.		**0.209**	0.310	0.370	0.399	0.338	1.099

Time Series Anomaly Detection

- Given any time series slice $x_1, x_2, ..., x_t$, the task of time series anomaly detection is to determine whether the last point x_t is an anomaly.
- The anomaly score is redefined based on the representations computed from masked and unmasked inputs during the inference stage.
- The TS2Vec model forwards twice for an input: first with the last observation x_t masked, and second with no mask applied.
- Representations of the last timestamp from these two forwards are denoted as r_t^u and r_t^m respectively.
- L_1 distance between r_t^u and r_t^m is used to measure the anomaly score: $\alpha_t = \|r_t^u - r_t^m\|_1$.
- A local average of the preceding Z points is taken to adjust the anomaly score, and a standardized score α_t^{adj} is calculated.
- A timestamp t is predicted as an anomaly if $\alpha_t^{\text{adj}} > \mu + \beta\sigma$, where μ and σ are the mean and standard deviation of the historical scores, and β is a hyperparameter.

Time Series Forecasting

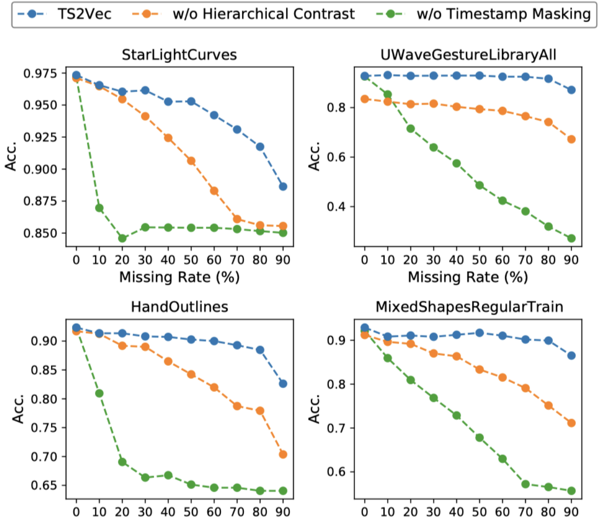
	Yahoo			KPI		
	F_1	Prec.	Rec.	F_1	Prec.	Rec.
SPOT	0.338	0.269	0.454	0.217	0.786	0.126
DSPOT	0.316	0.241	0.458	0.521	0.623	0.447
DONUT	0.026	0.013	0.825	0.347	0.371	0.326
SR	0.563	0.451	0.747	0.622	0.647	0.598
TS2Vec	**0.745**	0.729	0.762	**0.677**	0.929	0.533
Cold-start:						
FFT	0.291	0.202	0.517	0.538	0.478	0.615
Twitter-AD	0.245	0.166	0.462	0.330	0.411	0.276
Luminol	0.388	0.254	0.818	0.417	0.306	0.650
SR	0.529	0.404	0.765	0.666	0.637	0.697
TS2Vec†	**0.726**	0.692	0.763	**0.676**	0.907	0.540

Table 4: Univariate time series anomaly detection results.

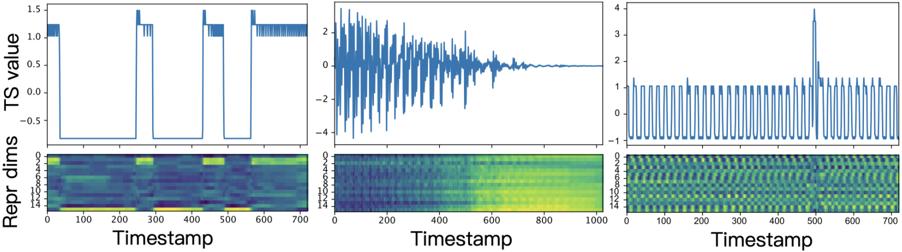
Ablation Study

	Avg. Accuracy
TS2Vec	**0.829**
w/o Temporal Contrast	0.819 (-1.0%)
w/o Instance Contrast	0.824 (-0.5%)
w/o Hierarchical Contrast	0.812 (-1.7%)
w/o Random Cropping	0.808 (-2.1%)
w/o Timestamp Masking	0.820 (-0.9%)
w/o Input Projection Layer	0.817 (-1.2%)
Positive Pair Selection	
Contextual Consistency	
$\rightarrow$ Temporal Consistency	0.807 (-2.2%)
$\rightarrow$ Subseries Consistency	0.780 (-4.9%)
Augmentations	
+ Jitter	0.814 (-1.5%)
+ Scaling	0.814 (-1.5%)
+ Permutation	0.796 (-3.3%)

Robustness to Missing Data



Visualized Explanation



(a) ScreenType. (b) Phoneme. (c) RefrigerationDevices.

Section 7: Summary

- Propose a unified framework that learns contextual representations for arbitrary sub-series at various semantic levels
- Purpose hierarchical contrasting method in both instance-wise and temporal dimensions to capture multi-scale contextual information.
- Propose contextual consistency for positive pair selection.
- Perform well in downstream tasks, efficient on training, and robust to missing data