ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

Dynamics-aware Representation Learning via Multivariate Time Series Transformers

Michael Potter[1*], İlkay Yıldız Potter [2*], Octavia Camps[3] and Mario Sznaier [3]

1- Naval Surface Warfare Center Corona, Norco, CA, USA
michael.l.potter40.civ@us.navy.mil
2- BioSensics LLC, Newton, MA, USA
ilkay.yildiz@biosensics.com
3- Northeastern University, Electrical and Computer Engineering, Boston, MA, USA
camps@ece.neu.edu, msznaier@ece.neu.edu

Abstract. We propose a novel multivariate time series autoencoder, which produces interpretable linear-dynamical latent features that govern the predictions for several downstream tasks. To this end, we combine a transformer autoencoder with a dynamical atoms-based autoencoder to mimic Koopman operators in the latent space. We demonstrate that our approach significantly outperforms deep Koopman operator learning baselines for time series forecasting on chaotic systems such as the lorenz Attractor. Furthermore, the dynamics-aware representations, combined with a transformer classifier, lead to state-of-the-art classification accuracy on benchmark multivariate time series datasets. Our code is publicly available at `https://github.com/mlpotter/T-DYAN-T`.

1 Introduction

Time series data prevalently emerge in many domains, including, e.g., healthcare, finance and autonomous driving. Unlike the domination of deep learning in computer vision and natural language processing, linear models such as HIVE-COTE [1], TS-CHIEF [2] and ROCKET [3] have been favorable to deep learning for time series analysis. Virtually only transformer models have significantly outperformed shallow models in supervised learning over multivariate time series (MVTS) benchmarks [4]. Particularly, pre-training overparametrized deep models such as transformers via autoencoder-based unsupervised representation learning has led to the breakthrough in downstream supervised tasks.

Despite the recent success of deep learning in several applications, the unique nature of real-life MVTS data that capture the dynamic evolution of synchronous variables has not yet been systematically imposed into deep models. To this end, there has been promising advances towards dynamics-aware representation learning from time-series. Particularly, recent deep autoencoder approaches [5, 6] aim to estimate data-driven representations of Koopman eigenfunctions in the latent space, which provide intrinsic coordinates that globally linearize dynamics. Nevertheless, these works have been limited to reconstruction and forecasting on well-established physical models, including, e.g., linear and nonlinear oscillators and the chaotic lorenz attractor.

The overall dominance of linear models in downstream tasks, along with the promising advances in dynamics-aware deep representation learning on physical models, motivate an interpretable and generalizable deep representation learning method to reinforce the presence of deep learning in MVTS analysis. We make the following contributions:

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

- To the best of our knowledge, our main contribution is to propose the first dynamics-aware representation learning framework via MVTS transformers that consolidate predictions in *both* reconstruction and forecasting, as well as downstream supervised tasks such as classification.
- We employ a transformer autoencoder [7], where latent features are augmented with linear dynamics through a dynamical atoms-based autoencoder. Compared to Koopman operators, our formulation relaxes latent space assumptions, such as invariant eigenfunction spaces and full-state observability of features.
- Our method significantly improves forecasting over two unsupervised autoencoder baselines that approximate finite Koopman operators. Furthermore, training a transformer classifier on the features learned by the transformer autoencoder, we outperform several MVTS classification baselines.

2 Problem Formulation

We consider a dataset of N MVTS samples, each comprising M features and $T+n$ points. Formally, we denote each MVTS by $\boldsymbol{X}^{(i)} \in \mathbb{R}^{T+n\times M}$, for $i \in \{1,\ldots,N\}$. Our aim is to learn dynamics-aware latent representations from MVTS data. To this end, we employ a transformer autoencoder [7], where latent feature learning is augmented with linear dynamics through a Dynamical Atoms-Based Network.

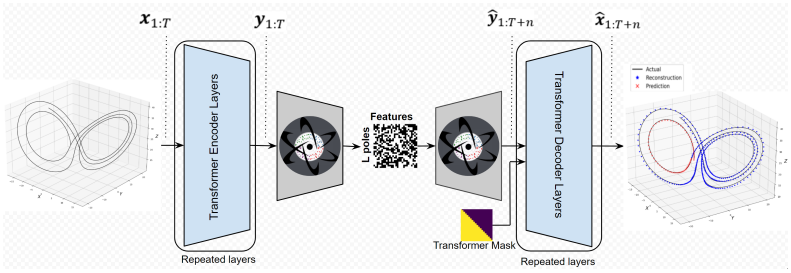


Figure 1: Our proposed transformer autoencoder

Transformer Autoencoder. Our transformer autoencoder architecture contains a transformer encoder followed by a transformer decoder [7]. The encoder network $\boldsymbol{\Phi}(\cdot;\boldsymbol{W}_{\boldsymbol{\Phi}})$ receives an MVTS sample $\boldsymbol{X}^{(i)}_{1:T}$ and extracts latent features $\boldsymbol{Y}^{(i)}_{1:T} \in \mathbb{R}^{T\times P}$. The decoder network $\boldsymbol{\Theta}(\cdot;\boldsymbol{W}_{\boldsymbol{\Theta}})$ receives the extracted latent features $\hat{\boldsymbol{Y}}^{(i)}_{1:T+n} \in \mathbb{R}^{T+n\times M}$ and estimates $\hat{\boldsymbol{X}}^{(i)}_{1:T+n} \in \mathbb{R}^{T+n\times M}$. The first T rows comprise the reconstruction from the T time points in the input sample $\boldsymbol{X}^{(i)}$, while the next n rows comprise the forecasted time points. To impose linear dynamics in the latent space of our transformer autoencoder, we employ the Dynamical Atoms-Based Network (DYAN) [8].

DYAN. DYAN [8] is an unsupervised autoencoder network that captures temporal dynamics via Linear Time Invariant (LTI) systems. Formally, DYAN learns a structured dictionary $\boldsymbol{D}_{1:T} \in \mathbb{R}^{T\times L+1}$ to encode latent sequences $\boldsymbol{Y}_{1:T}$ as a weighted summation of L low-order LTI systems. These systems are governed by poles $p_i = \rho_i e^{j\phi_i}$, $i \in$

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

$\{1, \ldots, L\}$, with magnitude ρ_i and phase ϕ_i. To combat the complex poles, DYAN employs a dictionary with real and imaginary parts of the poles in the first quadrant, of their conjugates and of their symmetries in other quadrants. As a result, each column (atom) of the dictionary $\boldsymbol{D}_{1:T}$ is the impulse response of a low order LTI system, where the first column corresponds to constant signals:

$$\boldsymbol{D}_{1:T} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & \rho_1 cos(\phi_1) & \rho_1 sin(\phi_1) & \cdots & -\rho_L sin(\phi_L) \\ 1 & \rho_1^2 cos(2\phi_1) & \rho_1^2 sin(2\phi_1) & \cdots & -\rho_L^2 sin(2\phi_L) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \rho_1^T cos(T\phi_1) & \rho_1^T sin(T\phi_1) & \cdots & -\rho_L^T sin(T\phi_L) \end{bmatrix}. \tag{1}$$

DYAN learns the low-order LTI systems by:

$$\boldsymbol{C}^* = \min_{\boldsymbol{C}} \sum_{i=1}^{N} \left\| \Phi(\boldsymbol{X}_{1:T}^{(i)}; \boldsymbol{W}_{\boldsymbol{\Phi}}) - \boldsymbol{D}_{1:T}^{(i)} \boldsymbol{C} \right\|_2^2 + \lambda_1 \left\| \boldsymbol{C} \right\|_1. \tag{2}$$

The coefficient matrix $\boldsymbol{C} \in \mathbb{R}^{L+1 \times P}$ reconstructs the latent features and is designed to be a sparse matrix due to the ℓ_1 penalty; this design allows $\boldsymbol{C}$ to select as few low order LTI systems as possible. More importantly, due to its proximal form, (2) can be solved by the efficient fast iterative shrinkage-thresholding (FISTA) algorithm. To extend latent features to future time points, we add rows to $\boldsymbol{D}_{1:T}$ (1) for $T+1 : T+n$. Having extended to $\boldsymbol{D}_{1:T+n} \in \mathbb{R}^{T+n \times L+1}$ as above, we estimate the reconstructed and forecasted latent features using the same optimal $\boldsymbol{C}^*$: $\hat{\boldsymbol{Y}}_{1:T+n} = \boldsymbol{D}_{1:T+n} \boldsymbol{C}^*$.

Dynamics-aware Representation Learning via Transformer Autoencoder. Putting everything together, we denote our unsupervised dynamics-aware MVTS representation learning architecture, depicted in Figure 1. Our approach involves a transformer autoencoder, in which DYAN receives $\boldsymbol{Y}_{1:T}^{(i)}$ and produces the reconstructed and forecasted latent features $\hat{\boldsymbol{Y}}_{1:T+n}^{(i)}$ by enforcing linear dynamics in the latent space.

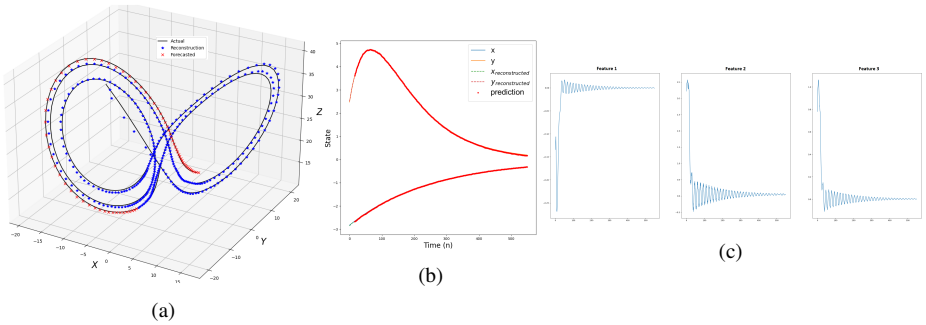
The end-to-end loss function (3) to train the transformer autoencoder has three ℓ_2 norm objectives: enforced linear dynamics via (2), reconstruction and forecasting:

$$\min_{\boldsymbol{W}_{\boldsymbol{\Phi}}, \boldsymbol{W}_{\boldsymbol{\Theta}}, \boldsymbol{W}_{\boldsymbol{p}}} \sum_{i=1}^{N} \left\| \boldsymbol{X}_{1:T}^{(i)} - \hat{\boldsymbol{X}}_{1:T}^{(i)} \right\|_2^2 + \lambda_2 \left\| \boldsymbol{X}_{T+1:T+n}^{(i)} - \hat{\boldsymbol{X}}_{T+1:T+n}^{(i)} \right\|_2^2, \tag{3}$$

where $\boldsymbol{W}_{\boldsymbol{p}}$ contain magnitudes and phases of DYAN poles, and the reconstructed and forecasted time points are estimated by the decoder as $\hat{\boldsymbol{X}}_{1:T+n} = \Theta(\boldsymbol{D}_{1:T+n} \boldsymbol{C}^*; \boldsymbol{W}_{\boldsymbol{\Theta}})$.

Extension to Downstream Classification. To illustrate the benefit of dynamics-aware latent representations in downstream tasks, we employ the learned latent features $\hat{\boldsymbol{Y}}_{1:T+n}^{(i)}$ as inputs for downstream classification. To do so, we combine a transformer encoder with a fully-connected classification layer. For each MVTS sample i, the transformer classifier receives latent features $\hat{\boldsymbol{Y}}_{1:T+n}^{(i)}$ and predicts the corresponding classification label. We employ the same transformer encoder architecture as 2, except for positional encoding. Note that the applicability of the dynamics-aware latent features is not limited to this illustrative task and can be extended to, e.g., regression.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.



(a) (b) (c)

Figure 2: (a) Reconstruction and forecasting on a lorenz test sample, (b-c) Reconstruction and forecasting, and latent features of a Carleman test sample

Model	Transformer + DYAN	Fully-connected + DYAN	Lusch et al. (2018)	Geneva et al. (2022)
MSE	**0.273**	0.348	0.3850	1.124

Table 1: Mean-squared forecasting error (MSE) over all test samples from the lorenz Attractor

3 Experiments

3.1 Datasets

lorenz Attractor. We consider the lorenz Chaotic system:

$$\dot{x}_1 = \sigma(x_2 - x_1); \dot{x}_2 = x_1(\rho - x_3); \dot{x}_3 = x_1 x_2 - \beta x_3; \tag{4}$$

with parameters $\sigma = 28$, $\rho = 10$, and $\beta = \frac{8}{3}$. Using MATLAB ODE45, we generate 15000 and 1000 trajectories of 400 points for training and testing, respectively. We use $T = 328$ points for reconstruction and the remaining $n = 72$ points for forecasting.
Carleman Nonlinear Model. We consider the simple nonlinear differential equation:

$$\dot{x}_1 = \mu x_1; \dot{x}_2 = \lambda(x_2 - x_1^2); \tag{5}$$

with parameters $\mu = -0.2$ and $\lambda = -0.8$. Using MATLAB ODE45, we generate 15000 and 1000 trajectories of 550 points for training and testing, respectively. We use $T = 32$ points for reconstruction and the remaining $n = 534$ points for forecasting.
Downstream Classification. We employ 4 benchmark MVTS classification datasets from the UEA Time Series Classification Archive [9]. For data partitioning, we use the training and test splits provided by the archive.

3.2 Competing Methods

Unsupervised Dynamics-Aware Representation Learning. We compare the forecasting performance of our representation learning approach over the test samples generated from lorenz Attractor with two unsupervised fully-connected autoencoder models that approximate finite Koopman operators for linear dynamics in the latent space: an autoencoder with a dynamically learnable Koopman operator representing continuous

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Dataset	Ours			TST		Rocket	XGBoost	LSTM	CNN	DTW_D
	Max	Mean	p-value	Max	Mean					
FaceDetection	0.693	**0.689±0.004**	**0.02**	0.688	0.683±0.004	0.647	0.633	0.577	0.528	0.529
ArabicDigits	0.995	**0.993±0.001**	**0.02**	0.992	0.992±0.000	0.712	0.696	0.319	0.956	0.963
PEMS-SF	0.861	0.851±0.011	**0.05**	0.884	0.825±0.031	0.751	**0.983**	0.399	0.688	0.711
Heartbeat	0.785	**0.778±0.013**	0.14	0.781	0.770±0.009	0.756	0.732	0.722	0.756	0.717
Mean Accuracy	**0.828**			0.818		0.717	0.761	0.504	0.732	0.730
Mean Ranking	**1.25**			2.25		3.875	4	6.25	5.125	5.25

Table 2: Average classification accuracy and ranking of each method. For transformers, we report maximum and standard deviation over 5 repetitions, and p-value of the improvement by ours.

eigenvalue spectra [5] and an autoencoder enforcing the Koopman operator to be a band matrix [6]. Moreover, we implement a baseline for our approach with a fully-connected autoencoder comprising 2 encoder and 2 decoder layers, instead of a transformer. We report mean-squared errors for all competing methods.

Downstream Classification. Following the recent advances in MVTS classification [4], we compare the downstream classification accuracy of our approach over the test samples with several recent baseline methods: a dilation convolutional neural network (CNN) [10], one nearest neighbour classifier with dimension dependent dynamic time warping similarity (DTW_D), ROCKET, XGBoost [11], a stacked long-short term memory (LSTM) network, and an MVTS transformer (TST) [4]. For transformer models, we report average, maximum and standard deviation of the accuracy over 5 training repetitions, as well as the p-value of the performance improvement by our approach under the one-sided t-test. We implement the TST classifier baseline following the best hyperparameters reported in the recent literature for each dataset [4].

3.3 Performance on Reconstruction and Forecasting

Table 1 shows the forecasting errors of deep Koopman learning autoencoder models against our dynamics-aware representation learning approach. Despite the underlying chaotic dynamics of lorenz and the extensive forecasting window of Carleman, our method attains a reconstruction error of 0.217 (0.0004) and a forecasting error of 0.273 (0.0005) over the test samples from lorenz (Carleman). Crucially, we significantly outperform all unsupervised representation learning baselines, while both transformer and fully-connected versions of our approach via linear latent dynamics enforced by DYAN outperform the deep Koopman learning competitors, illustrating the benefit of dynamical modeling via DYAN and relaxing the latent space assumptions.

Examples of reconstructed and forecasted time points on a test sample from the lorenz (Carleman) model are shown in Figure 2a (2b). Our method can not only learn underlying dynamics via reconstruction, but also generalize as well over forecasted points, which is fully governed by extending the DYAN model in the latent space. We further visualize the latent space trajectories on the same Carleman test sample in Figure 2c. The ground-truth closed linear system expansion of Carleman equations appear as exponentially decaying sinusoidal functions, while the estimated long-term linear dynamics are also exponentially decreasing sinusoidals after the sharp overshoot. This observation further illustrates that the learned latent features can successfully capture the ground-truth dynamics that govern reconstruction and forecasting.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

3.4 Performance on Downstream Classification

Table 2 show the downstream classification accuracy of our transformer classifier trained on transformer autoencoder latent features vs. MVTS classification baselines evaluated over 4 benchmark datasets. Our approach attains the highest average accuracy and performance ranking among all baselines, while outperforming the TST baseline with a statistically significant margin over 3 datasets. Overall, our approach not only leads to striking forecasting over chaotic physical systems, but also produces dynamics-aware representations that do not sacrifice from and even improve against downstream classification compared to several state-of-the-art methods.

4 Conclusion

We developed a novel unsupervised dynamics-aware representation learning framework that is competitive with benchmark MVTS methods using autoencoders, time series forecasting, and feature extraction for time series classification. Future work can involve discovering the latent dimension size adaptively by uncovering the nullspace trajectories of DYAN reconstructions. Moreover, dynamics-aware latent features can be combined with explainable machine learning to underline interpretable representation learning.

References

[1] Jason Lines, Sarah Taylor, and Anthony Bagnall, "Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 5, 2018.

[2] Ahmed Shifaz, Charlotte Pelletier, François Petitjean, and Geoffrey I Webb, "Ts-chief: a scalable and accurate forest algorithm for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 3, pp. 742–775, 2020.

[3] Angus Dempster, François Petitjean, and Geoffrey I Webb, "ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.

[4] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[5] Bethany Lusch, J Nathan Kutz, and Steven L Brunton, "Deep learning for universal linear embeddings of nonlinear dynamics," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[6] Nicholas Geneva and Nicholas Zabaras, "Transformers for modeling physical systems," *Neural Networks*, vol. 146, pp. 272–289, 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] WenQian Liu, Abhishek Sharma, Octavia I. Camps, and Mario Sznaier, "DYAN: A dynamical atoms network for video prediction," *CoRR*, vol. abs/1803.07201, 2018.

[9] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data mining and knowledge discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[10] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi, "Unsupervised scalable representation learning for multivariate time series," *Advances in neural information processing systems*, vol. 32, 2019.

[11] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.