INTERPRETABLE REPRESENTATION SPACES FOR TIME SERIES

A PREPRINT

ABSTRACT

In some machine learning applications it is essential for models to provide interpretable predictions. In domains such as computer vision and natural language processing, there are algorithms for disentanglement of hidden states. Disentangled representations of hidden states contain human-interpretable information about objects, making it easier to describe model predictions using only these representations.

In this paper, we introduce an efficient pipeline for generating interpretable representations from multidimensional time series data. Unlike other disentanglement algorithms for sequential data, the proposed method creates a single constant-dimensional representation for the entire sequence of measurements. This approach enables global changes to time series to be made in a controlled manner, offering greater flexibility and interpretability.

Keywords Few keywords here

1 Introduction

In many machine learning applications, especially those with real-world impact, interpretability is crucial for model predictions. Interpretability often achieved by constructing interpretable representations of the hidden states within deep learning models. Also, some generative modeling approaches exploit variations in a limited set of parameters in an interpretable latent space to facilitate tasks such as controlled generation.

In natural language processing (NLP), interpretable representation spaces are often constructed using techniques like sparse autoencoders (SAEs) Trenton Bricken et al. [2023]Lieberum et al. [2024], which help to disentangle hidden states. Variational Autoencoders (VAEs) have been successfully employed in supervised tasks where predefined interpretability terms are available. Additionally, various methods (e.g., METHOD NAME) are designed for controlled generation, where interpretable latent dimensions guide output characteristics.

In this work, we propose a novel modification of the VAE framework tailored to time series data, specifically designed to decompose complex, multidimensional time series into independent components of trend and seasonality. We demonstrate the effectiveness of our approach in the context of human-skeleton motion, showing that the learned interpretable representations allow for plausible, synthetic generation of motion sequences, thereby advancing interpretable time series modeling.

2 Related Works

In generative models, disentangled latnent space can be obtaining by designing the specific architecture of neural network Kingma [2013]Higgins et al. [2017] or optimizing additional loss functionsChen et al. [2018]Balabin et al. [2023]. While the latter approach can admit supervised learning Paige et al. [2017], the most challenging but practical approach is unsupervisedDenton et al. [2017] learning of disentangled representations since the underlying factors of variation are typically unknown for real data.

Variational Autoencoders (VAEs) Kingma [2013] approximate the latent variable distribution by aligning it with a predefined target distribution, such as a normal distribution with a unit covariance matrix. However, the original VAE architecture often fails to achieve accurate approximation of the target distribution $p(z)$, which limits its ability to produce disentangled representations. To address this limitation, various modifications of the architecture have been

proposed. For instance, the β-VAE Higgins et al. [2017] introduces an increased weight for the KL divergence between the latent variable distribution and the target distribution, thereby promoting better disentanglement at the cost of reconstruction fidelity. Further advancements include architectures like β-TCVAE Chen et al. [2018] and Factor-VAE Kim and Mnih [2018], which explicitly optimize for total correlation within the latent variables. These approaches aim to enforce independence among the latent factors, thus achieving a more robust disentangled representation.

There are approaches based on the manifold hypothesis Goodfellow [2016] which posits that data points are concentrated in a vicinity of a low-dimensional manifold. For disentangled representations, it is crucial that the manifold has a specific property, namely, small topological dissimilarity between a point cloud given by a batch of data points and another point cloud obtained via the symmetry group(oid) action shift along a latent space axis. In the Balabin et al. [2023] was proposed method for unsupervised learning of disentangled representations via adding to a VAE-type loss the topological objective.

TimeVAE Desai et al. [2021] introduces a method for supervised learning of interpretable representation spaces for time series. This approach is based on the decomposition of time series into a sum of trend and seasonal components, enabling the model to capture distinct patterns within the data and improve the interpretability of the latent representations.

3 Background

3.1 Variational Autoencoder

The Variational Autoencoder (VAE) Kingma [2013] is a generative model that encodes an object x_n into a set of parameters of the posterior distribution $q_\phi(z|x_n)$, represented by an encoder with parameters ϕ. Then it samples a latent representation from this distribution and decodes it into the distribution $p_\theta(x_n|z)$, represented by a decoder with parameters θ. The prior distribution for the latent variables is denoted as p(z). In this work, we consider the factorized Gaussian prior $p(z) = \mathcal{N}(z \mid 0, I)$, and the variational posterior for an observation is also assumed to be a factorized Gaussian distribution with the mean and variance produced by the encoder. The standard VAE model is trained by minimizing the negative Evidence Lower Bound (ELBO) averaged over the empirical distribution:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{Rec}} \tag{1}$$

$$\mathcal{L}_{\text{KL}} = -\frac{1}{N}\sum_{n=1}^{N} \text{KL}(q_\phi(z \mid x_n) \parallel p(z)) \tag{2}$$

$$\mathcal{L}_{\text{Rec}} = -\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q(z)} \log p_\theta(x_n \mid z) \tag{3}$$

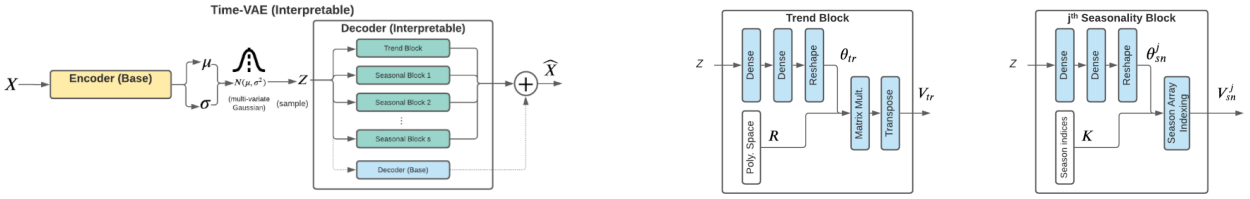
3.2 TimeVAE

The TimeVAE Desai et al. [2021] architecture (fig. 1) is a generative model for time series that encodes an object X into latent representation z with posterior distribution $q_\phi(z \mid x)$, represented by the encoder E with parameters ϕ. The reconstructed time series $\hat{X}$ obtained from decoder D with parameters ψ.

A time series is represented as the sum of trend and seasonality components, $X = X_{\text{season}} + X_{\text{trend}}$. An additive assumption is introduced for the latent representations: $z = z_{\text{season}} + z_{\text{trend}}$. The proposed architecture employs a single-stage encoder E and a two-stage decoder D, as follows:

1. Trend component $\hat{X}_{\text{trend}}$ is estimated from latent representation z. The latent representations of trend component z_{trend} is estimated from generated $\hat{X}_{\text{trend}}$ as $z_{\text{trend}} = E(\hat{X}_{\text{trend}})$
2. The season latent representation is computed as $z_{\text{season}} = z - z_{\text{trend}}$. The season component $\hat{X}_{\text{season}}$ is estimated from z_{season}
3. The final reconstructed time series is obtained as sum $\hat{X} = \hat{X}_{\text{season}} + \hat{X}_{\text{trend}}$

The optimization is described by (1)-(3).



(a) TimeVAE architecture (b) TimeVAE blocks architecture

Figure 1: TimeVAE architecture

3.3 DCI metric

The DCI Metric (Disentanglement, Completeness, Informativeness) Eastwood and Williams [2018] provides a quantitative framework for evaluating the quality of representations in machine learning models. This metric decomposes the evaluation into three complementary aspects: **disentanglement** (4), which measures the degree to which a representation factorises or disentangles the underlying factors of variation, with each variable (or dimension) capturing at most one generative factor; **completeness** (5), which quantifies the degree to which each underlying factor is captured by a single code variable; and **informativeness** which reflects the predictive capacity of the latent representations for downstream tasks. By jointly assessing these properties, the DCI metric offers a comprehensive perspective on the effectiveness and interpretability of learned representations.

$$P_{i,k} = \frac{R_{i,k}}{\sum_k R_{i,k}}$$

$$D = \frac{1}{h}\sum_{i}^{h}\left(1 + \sum_{k}^{K} P_{i,k}\log P_{i,k}\right) \tag{4}$$

$$C = \frac{1}{K}\sum_{k}^{K}\left(1 + \sum_{i}^{h} P_{i,k}\log P_{i,k}\right) \tag{5}$$

4 Method

This work introduces a modification of the architecture proposed in TimeVAE Desai et al. [2021]. This work proposes the following modifications to the described architecture. First, a conditional variational autoencoder (CVAE) is employed:

$$p_\phi(z \mid x, \theta) = \mathcal{N}(z \mid \mu(\theta), \Sigma(\theta)) \tag{6}$$

where $\mu(\theta)$ and $\Sigma(\theta)$ are neural networks.

Second, it is assumed that the latent representation z is partitioned into $K+1$ disjoint blocks J_k, each corresponding to different interpretable components θ_k. Independent models $\mu_{\phi_k}(\theta)$ and $\Sigma_{\phi_k}(\theta)$ are proposed to estimate the distributions for these blocks:

$$\begin{aligned} p_\phi(z \mid x, \theta) &= \mathcal{N}(z \mid \mu(\theta), \Sigma(\theta)) = \\ &= \mathcal{N}(z_{J_0} \mid 0, I)\prod_k \mathcal{N}(z_{J_k} \mid \mu_{\phi_k}(\theta), \Sigma_{\phi_k}(\theta)) \end{aligned} \tag{7}$$

This approach aims to reduce dependencies between components associated with different interpretable concepts. Also this approach allows to easily add new interpretable components θ_k by adding new disjoint blocks J_k into latent representation z.

5 Experiments

We conduct a variety of experiments to demonstrate the capabilities of VAE models for constructing the interpretable representations of time series. Firstly, we evaluate the benefits of interpretable representations in discriminative tasks. Subsequently, we demonstrate the opportunity to controllable generation of time series by performing variations in interpretable latent space. Furthermore, we compare the TimeVAE architecture with proposed architecture.

5.1 Datasets

The following datasets were used to conduct experiments.

The **Sine** synthetic dataset consists of realizations of deterministic function on a uniform grid. The following algorithm is used to the generation of proposed functions. First, we sample parameters of functions:

$$\alpha_i \sim \mathbb{U}(0.3,\ 5), \quad \psi_i \sim \mathbb{U}(0,\ 2\pi), \quad \phi_i \sim \mathbb{U}(0.1,\ \frac{\pi}{2}),$$

Subsequently, we compute projections on the uniform grid:

$$f_i(t) = \alpha_i cos(\phi_i \cdot (t + \psi_i)), \ \ t = \overline{1, 50}$$

Finally, the time series realization X_j computes as sum of few built functions:

$$X_j(t) = \sum_{i \in I_j} f_i(t), \quad |I_j| = K$$

The set of α_i, ϕ_i, ψ_j used to generate the X_j could be used as interpretable terms.

The **StockV** represents a standard set of financial time series introduced by the authors [REFERENCE]. A distinctive feature of this dataset is the significant variability and differences in the mean values of its components. After normalization, the components exhibit magnitudes on the order of 10^{-1}, with the first three components showing short-term deviations on the order of 10^{-4}, the next two components on the order of 10^{-3}, and the sixth component—representing the total trading volume — on the order of 10^{-2}.

No interpretable representations are explicitly defined for this task. However, it is possible to introduce certain concepts, such as the direction of short-term trends, trading seasons, or to employ technical analysis tools to derive domain-specific concepts. Nonetheless, in this study, the dataset is used solely to evaluate the performance of autoencoders on noisy data.

For this study, the **UPENN** Action dataset [REFERENCE] is employed as an example of real-world time series with distinguishable concepts. This dataset comprises $2\ 500$ videos and their corresponding annotations of human skeleton trajectories. The time series of 13 skeletal keypoints, defined by their x and y coordinates relative to the camera.

For interpretable concepts, the following categories were selected:

Human-performed actions: throwing and catching a baseball, bowling, golfing, pull-ups, push-ups, squats, and other physical exercises

Camera viewpoint relative to the person: frontal view, rear view, left-side view, and right-side view

5.2 Comparison of models

This section presents a comparsion between the TimeVAE architecture and the proposed architecture. During the experiment, each model was trained 15 times. Table 1 summarizes the reconstruction quality, evaluating the impact of the proposed modifications when applied to: the encoder, the decoder, and both the encoder and decoder. Having shown that the proposed method can achieve better reconstruction quality than base TimeVAE architecture.

	TV-TV	TV-PR	PR-TV	TV-TV
UPENN	0.054 ± 0.013	0.050 ± 0.012	0.052 ± 0.013	0.049 ± 0.012
Sine	0.073 ± 0.020	0.061 ± 0.018	0.064 ± 0.017	0.054 ± 0.015
Stockv	0.011 ± 0.012	0.010 ± 0.010	0.011 ± 0.010	0.010 ± 0.008

Table 1: Reconstruction RMSE. TV — the TimeVAE part, PR — the proposed part

Table 2 compares the quality of the constructed interpretable representations. The proposed architecture achieves a higher modularity score, which aligns with the assumptions underlying the model design. Specifically, introducing factor independence not only in the distribution $p(\tilde{z} \mid X)$ but also within the model implementing it enables a more effective decomposition into independent and interpretable components.

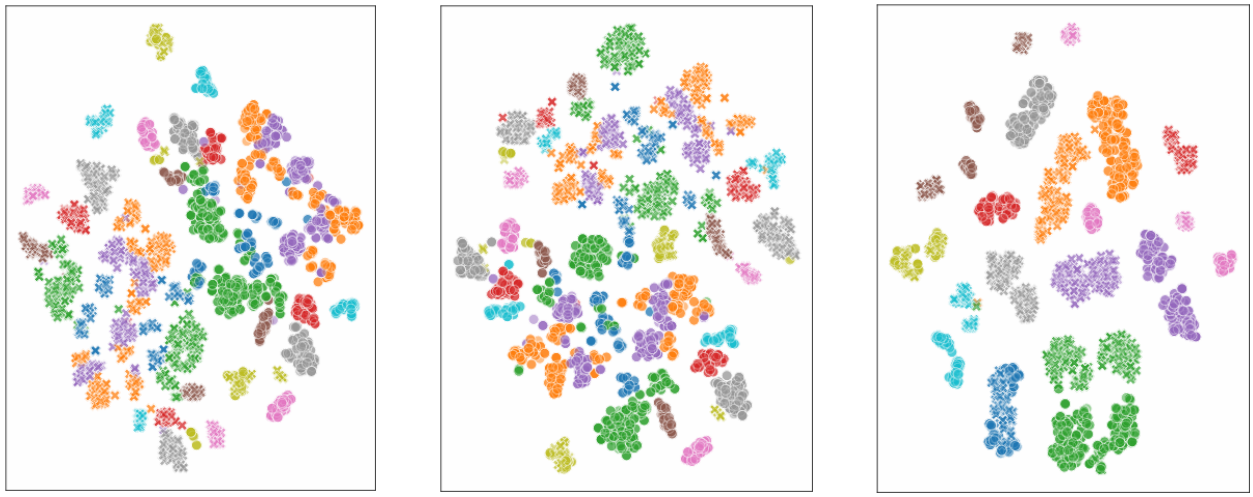
5.3 Use in discriminative tasks

This section demonstrates the potential use of interpretable predictions for constructing discriminative models. Figure fig. 2 depicts the t-SNE[REFERENCE] visualizations for the original time series, their latent representations, and

	TimeVAE	Proposed
Disentanglement	0.21	0.91
Completeness	0.07	0.26
Informativeness	0.94	1.00

Table 2: DCI metric on UPENN dataset

the interpretable representations. This experiment serves as a preliminary exploration, illustrating the hypothetical feasibility of constructing interpretable discriminative models based on the derived interpretable representations of time series.



(a) TSNE of raw time series

(b) TSNE of latent representations

(c) TSNE of interpretable representations

Figure 2: TSNE of time series representation from the UPENN dataset. Color describes the interpretable terms

5.4 Use in generation tasks

The model takes pairs of concepts as input to generate corresponding time series. Figures fig. 3, fig. 4, fig. 5, fig. 6 show examples of the generated series. All queries produced high-quality time series that align with the concepts embedded in them. This can be verified by analyzing the motion direction. In the visualizations, the left side of the human skeleton is highlighted in blue, and the right side in red.

For example, in the illustrated in fig. 3 case of a bowling action with a rear view, the person swings the ball backward with their right hand and then propels it forward, away from the camera. These results demonstrate that the proposed model is not only highly interpretable but also capable of generating high-quality time series.

The only limitation observed pertains to the accuracy of the direction-related viewpoint concepts. Specifically, there is a slight deviation in the angle of the viewpoint, which stems from the training dataset. The data originates from recordings of professional matches, where camera operators often aim to capture the most aesthetically pleasing angles. As a result, many examples labeled as "rear view" or "front view" were actually recorded with slight deviations. Thus, this issue is inherently tied to the characteristics of the dataset.

6 Conclusion

This study introduces a methodology for obtaining efficient interpretable representations of multivariate time series. A significant improvement to the existing autoencoder model is proposed, incorporating the assumption of decomposing time series into independent components of trend and seasonality.

Interpretable representations were constructed for time series describing human skeleton motions. Based on these representations, the ability to generate novel, unique, and plausible time series was demonstrated.

References

Adly Templeton Trenton Bricken, Brian Chen Joshua Batson, Nicholas L Turner Adam Jermyn, Tom Conerly, Carson Denison Cem Anil, Robert Lasenby Amanda Askell, Yifan Wu, Nicholas Schiefer Shauna Kravec, Tim Maxwell, Alex Tamkin Nicholas Joseph, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. Disentanglement learning via topology. *arXiv preprint arXiv:2308.12696*, 2023.

Brooks Paige, Jan-Willem Van De Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, Philip Torr, et al. Learning disentangled representations with semi-supervised deep generative models. *Advances in neural information processing systems*, 30, 2017.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018.

Ian Goodfellow. Deep learning, 2016.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018.

A Generation

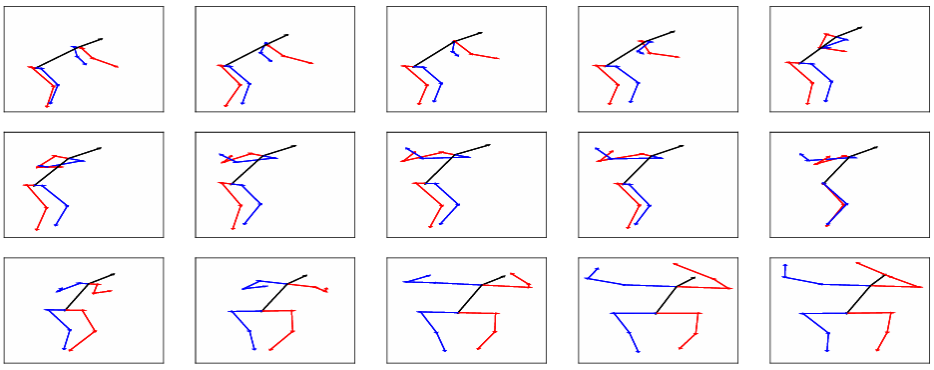


Figure 3: Golf, back view

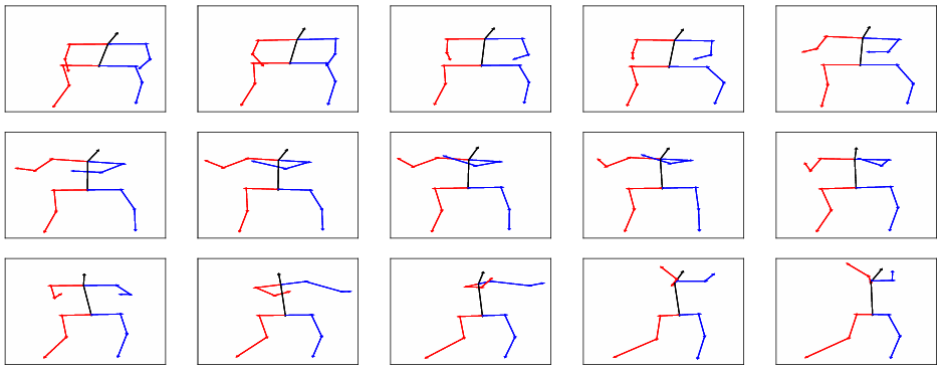


Figure 4: Golf, right view

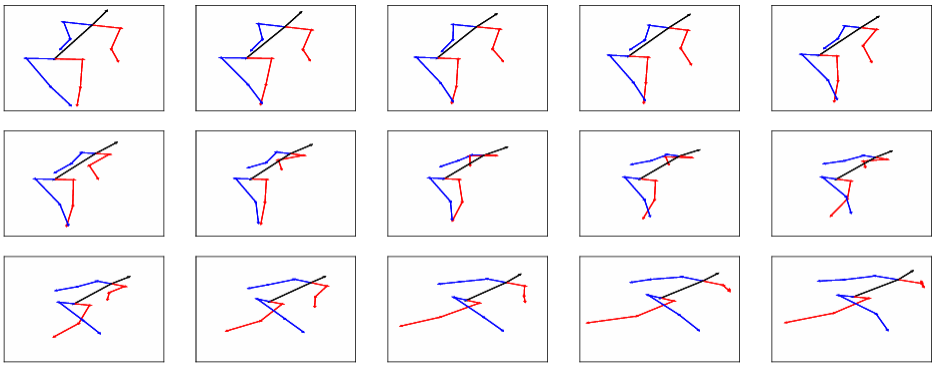


Figure 5: Bowling, back view

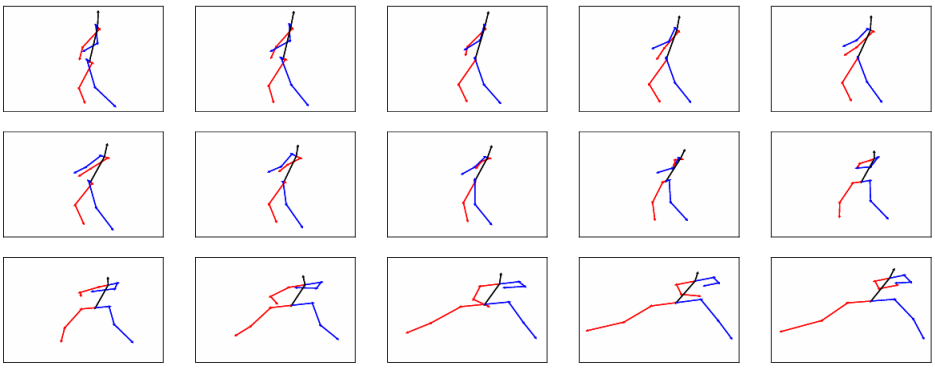


Figure 6: Bowling, front view