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:

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

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

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

3.2 TimeVAE

The TimeVAE Desai et al. [2021] architecture 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).

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

6 Conclusion

Список литературы

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.