Deep Variational Graph Convolutional Recurrent Network for Multivariate Time Series Anomaly Detection

Wenchao Chen [*1] **Long Tian** [*1] **Bo Chen** [1] **Liang Dai** [2] **Zhibin Duan** [1] **Mingyuan Zhou** [3]

Abstract

Anomaly detection within multivariate time series (MTS) is an essential task in both data mining and service quality management. Many recent works on anomaly detection focus on designing unsupervised probabilistic models to extract robust normal patterns of MTS. In this paper, we model channel dependency and stochasticity within MTS by developing an embedding-guided probabilistic generative network. We combine it with adaptive **V**ariational **G**raph **C**onvolutional **R**ecurrent **N**etwork (VGCRN) to model both spatial and temporal fine-grained correlations in MTS. To explore hierarchical latent representations, we further extend VGCRN into a deep variational network, which captures multilevel information at different layers and is robust to noisy time series. Moreover, we develop an upward-downward variational inference scheme that considers both forecasting-based and reconstruction-based losses, achieving an accurate posterior approximation of latent variables with better MTS representations. The experiments verify the superiority of the proposed method over the current state of the art.

1. Introduction

Multivariate time series (MTS) data are increasingly collected in various real-world systems, such as internet services (Su et al., 2019; 2021), content delivery networks (Dai et al., 2021), power plants (Len et al., 2007), and wearable devices (Djurdjanovic et al., 2003). Anomaly detection and diagnosis of MTS in information technology refers to identifying the abnormal status in certain time steps and pinpointing the root causes, and it has become an active topic in data science. Although traditional statistical-based anomaly detection methods (Chen et al., 2013; Siffer et al., 2017; Chandola et al., 2009) have been proved to be effective, they heavily rely on the prior knowledge and are unfriendly to complex MTS. Machine learning-based methods (Pang et al., 2020), which have drawn significant recent attention, can be classified into two primary categories, which are supervised (Liu et al., 2015; Shon & Moon, 2007; Yamada et al., 2013) and unsupervised (Malhotra et al., 2016; Hundman et al., 2018; Li et al., 2019; Dai et al., 2021; Su et al., 2021; Chen et al., 2019) methods. Due to the labor-intensive data annotation and the lack of anomaly instances in real-world scenarios, supervised methods tend to become impractical. Hence, the unsupervised anomaly detection has been widely studied in recent years. Among them, one line of the research mainly focuses on learning spatial characteristics in the multivariate metrics but ignores temporal dependencies across time steps (Zong et al., 2018; Audibert et al., 2020; Su et al., 2021). Another line of the work is recurrent neural networks (RNNs) based anomaly detection, modeling the temporal dependencies via recurrent structure (Malhotra et al., 2016; Li et al., 2019; Dai et al., 2021; Su et al., 2021). Besides, recent research also focuses on capturing inter-channel relationships of MTS with convolutional neural networks (CNNs) (Li et al., 2021; Zhang et al., 2019b; Chen et al., 2020) or graph neural networks (GNNs) (Deng & Hooi, 2021; Song et al., 2020; Bai et al., 2019), which have been proved to be effective in capturing normal patterns of MTS. However, all these methods ignore the stochasticity of MTS, thus failing to achieve the robust anomaly detection.

To consider the stochasticity of MTS, some probabilistic methods (Xu et al., 2018; Su et al., 2021), especially dynamical models (Dai et al., 2021; Li et al., 2021), are developed and exhibit promising performance. To further capture the inter-relationship, graph structures are also incorporated into the inference network of probabilistic dynamic models (Zhao et al., 2020). Despite the good performance that existing probabilistic anomaly detection methods claim, we notice two phenomena in MTS that they have difficulty dealing with effectively: ***1)*:** Anomalies in MTS are always

[*]Equal contribution [1]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. [2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. [3]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA. Correspondence to: Bo Chen <bchen@mail.xidian.edu.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

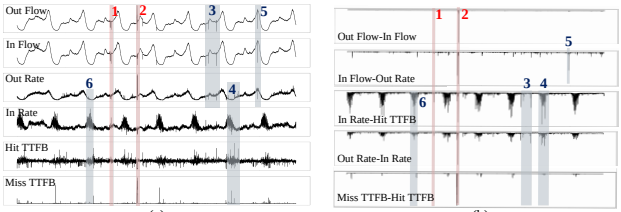


Figure 1. (a) 8 days MTS from the Distributed Network Dataset (DND), which are Key Performance Indicators (KPIs) collected from the distributed network; (b) Relations between channels in (a). Regions highlighted in red and blue represent the ground-truth anomaly and normal that is misjudged to anomaly by previous methods, respectively.

reflected by both the changes of temporal and relational characteristics, such as region 2 in Fig. 1, while many temporal fluctuations, such as regions 3 and 5, may easily lead to the misjudgment by probabilistic dynamic models. However, there is still no probabilistic method that is able to capture the relational information of MTS in the generative process and consider its stochasticity for the robust representation learning, which may restrict the models' capacity in capturing anomalies, while avoiding the misjudgment of fluctuations. ***2)*:** There are always some noisy time series in MTS, such as "Out rate," "In rate," and "Hit TTFB" in Fig. 1 (a), that can cause both temporal and relational fluctuations, thus easily leading to the misjudgment, as shown in region 4 in Fig. 1 (a) and (b). Most existing probabilistic methods are still shallow models (Sønderby et al., 2016; Zhou et al., 2016; Child, 2021; Guo et al., 2018; Zhang et al., 2018), which may have limited representational space, limiting their robustness to these noisy time series.

Retaining the advantages of existing probabilistic models in capturing the non-deterministic dynamics within MTS, while relaxing their constraints that can not model the non-deterministic relational information in generative process, we propose a novel **D**eep **V**ariational **G**raph **C**onvolutional **R**ecurrent **N**etwork (DVGCRN) that consists of a developed **D**eep **E**mbedding-guided **P**robabilistic generative **N**etwork (DEPN) to model hierarchical non-deterministic inter-relationships within MTS, **S**tacked **G**raph **C**onvolutional **R**ecurrent **N**etwork (SGCRN) to model multilevel temporal dependencies, and Gaussian-distributed channel embeddings to characterize the similarity and stochasticity of different channels, as illustrated in Fig. 6. For efficient inference, firstly, we construct an upward-downward autoencoding inference method. It combines the bottom-up likelihood and the up-bottom prior information of the parameters to perform the accurate posterior inference. Then, we introduce a joint optimization objective that combines the forecasting-based loss for SGCRN and the reconstruction-based loss for DEPN to ensure better time-series representation learning.

Our contributions can be summarized as follows:

- We propose VGCRN, a novel stochastic model consisting of the developed EPN and GCRN, for robust anomaly detection of MTS. It is capable of discovering both non-deterministic temporal dependencies and channel relationships within MTS.
- To explore hierarchical latent representations and long-term temporal dependencies, and reveal multilevel inter-relationships of MTS, we extend VGCRN to a deep stochastic model called DVGCRN.
- We develop an upward-downward inference scheme for accurate inference and integrating the advantages of both forecasting-based and reconstruction-based models by introducing a joint optimization objective.
- Experimental results on real-world and public datasets substantiate the superiority of DVGCRN as compared with the current state of the art.

2. Related work

Anomaly detection, an important task in MTS analysis, has attracted increasing concerns of operation engineers in recent years. Due to the superior ability in modeling complex functions, deep learning methods have dominated the studies of MTS. A majority of such works utilize recurrent structures to model the temporal dependencies in MTS. For example, a representative study called EncDec-AD (Malhotra et al., 2016) employs an LSTM-based encoder&decoder to capture the normal patterns of temporal dependencies for MTS and determine abnormal ones depending on reconstruction errors. Telemanom (Hundman et al., 2018) utilizes an LSTM (Sepp & Jurgen, 1997) to predict values of telemetry channels in the spacecraft and detect an anomaly based on residual errors between the predicted and observed values. To consider the stochasticity within MTS, OmniAnomaly (Su et al., 2019) and SDFVAE (Dai et al., 2021) present stochastic recurrent neural networks (SRNNs) to learn robust representations and detect anomalies by using the likelihood. Despite the effectiveness of existing dynamic methods in real-world scenarios, their ultimate potentials have been limited since they completely ignore the spatial correlations among channels within MTS.

To consider the inter-relationships of different channels within MTS, MSCRED (Zhang et al., 2019a) introduces a multi-scale convolutional recurrent encoder&decoder to learn spatial correlations and temporal characteristics in MTS and detect anomalies via residual signature matrices. InterFusion (Li et al., 2021) incorporates recurrent and convolutional structures into a unified framework to capture both temporal and inter-metric information. Recently, GNNs have gradually attracted more attentions in exploring the relationships. Thus, some GNN-based methods for anomaly detection have been developed (Deng & Hooi, 2021; Zhao et al., 2020). These works have been

proved to be effective in discovering normal patterns of MTS. However, these methods are still shallow models with a limited fitting power and fail to model stochastic inter relationships in a generative process.

3. Proposed model

In this section, we first define the problem solved in this paper. We then present VGCRN that integrates the GCRN and EPN into a well-defined unified framework for anomaly detection of MTS. Finally, we add a hierarchical generative procedure with multiple channel embeddings.

3.1. Problem definition

We define the n-th MTS as $\boldsymbol{x}_n = \{\boldsymbol{x}_{1,n}, \boldsymbol{x}_{2,n}, ..., \boldsymbol{x}_{T,n}\} \in \mathbb{R}^{T\times V}$, where $n = 1, ...N$ and N is the number of MTS. T is the duration of $\boldsymbol{x}_n$ and the observation at time t, $\boldsymbol{x}_{t,n} \in \mathbb{R}^V$, is a V dimensional vector and V denotes the number of time series. MTS anomaly detection is defined as a problem that determines whether an observation from a certain task and at a certain time is anomalous or not.

3.2. Channel embedding

To reflect the characteristics of different channels in MTS, similar with Deng & Hooi (2021), we first introduce an embedding vector for each channel in MTS as:

$$\boldsymbol{\alpha}_i^{(0)} \in \mathbb{R}^d,\ i \in \{1, 2, ..., V\},\ \boldsymbol{\alpha}^{(0)} = [\boldsymbol{\alpha}_1^{(0)}, ..., \boldsymbol{\alpha}_V^{(0)}]. \quad (1)$$

Then, the relationships between different channels can be indicated by the inner product of embeddings, and they are initialized randomly followed by being trained along with other parameters of the model. Besides, to capture inter-relationships of latent representations, we introduce the layer-wise embedding vectors as:

$$\boldsymbol{\alpha}_i^{(l)} \in \mathbb{R}^d, i \in \{1, 2, ..., K_l\}, l \in \{1, 2, ..., L\} \quad (2)$$

where K_l refers to the dimension of latent space of DEPN module. Moreover, to model stochastic of the inter-relationships, we define channel embeddings as Gaussian distributed vectors:

$$\boldsymbol{\alpha}_i^{(l)} = \mathcal{N}(\hat{\boldsymbol{\mu}}_i^{(l)}, \text{diag}(\hat{\boldsymbol{\sigma}}_i^{(l)})) \quad (3)$$

The embeddings are employed and optimized in both the DEPN and DGCRN modules of DVGCRN with both the reconstruction and prediction losses.

3.3. Variational graph convolutional recurrent network

3.3.1. Embedding-guided probabilistic generative network

Distinct from traditional probabilistic generative models that ignore the relations of different channels, we propose EPN that captures the inter-dependencies of MTS by introducing channel embeddings into generation process as:

$$\begin{aligned}
&\boldsymbol{W}_{z\mu}^x = \text{softmax}({\boldsymbol{\alpha}^{(0)}}^T \boldsymbol{\alpha}^{(1)}), \boldsymbol{W}_{h\mu}^x = \text{softmax}({\boldsymbol{\alpha}^{(0)}}^T \boldsymbol{\beta}) \\
&\boldsymbol{z}_{t,n} \sim \mathcal{N}\left(\boldsymbol{\mu}_{t,n}, \text{diag}\left(\boldsymbol{\sigma}_{t,n}\right)\right), \boldsymbol{\mu}_{t,n} = f\left(\boldsymbol{W}_{h,\mu}\boldsymbol{h}_{t-1,n}\right) \\
&\boldsymbol{\mu}_{t,n}^x = f\left(\boldsymbol{W}_{z\mu}^x \boldsymbol{z}_{t,n} + \boldsymbol{W}_{h\mu}^x \boldsymbol{h}_{t-1,n}\right) \\
&\boldsymbol{x}_{t,n} \sim \mathcal{N}\left(\boldsymbol{\mu}_{t,n}^x, \text{diag}\left(\boldsymbol{\sigma}_{t,n}^x\right)\right)
\end{aligned} \quad (4)$$

where $\boldsymbol{z}_{t,n} \in \mathbb{R}^K$ refers to the Gaussian distributed probabilistic latent variables, whose means and covariance parameters are $\boldsymbol{\mu}_{t,n}$ and $\boldsymbol{\sigma}_{t,n}$. $\boldsymbol{h}_{t,n} \in \mathbb{R}^{K'}$ denotes the deterministic latent states of the GCRN module, introduced in the next subsection. $f(\cdot)$ refers to the non-linear activation function and $\boldsymbol{\alpha}^{(l)} = [\boldsymbol{\alpha}_1^{(l)}, ..., \boldsymbol{\alpha}_{K_l}^{(l)}] \in \mathbb{R}^{d\times K_l}$. $\boldsymbol{\mu}_{t,n}^x$ and $\boldsymbol{\sigma}_{t,n}^x$ are the mean and variance parameters of $\boldsymbol{x}_{t,n}$, they are all trainable variables. The channel embeddings are incorporated into the generative process by defining the factor loading matrices $\boldsymbol{W}_{z\mu}^x$ as the inner product of them, thus to model the similarity of different channels and capture the complex inter-relationships between channels. $\boldsymbol{W}_{h,\mu}$ represents the connection matrix between $\boldsymbol{h}_{t-1,n}$ and $\boldsymbol{z}_{t,n}$, $\boldsymbol{W}_{h,\mu}^x$ represents the connection matrix between $\boldsymbol{h}_{t-1,n}$ and $\boldsymbol{x}_{t,n}$. $\beta \in \mathbb{R}^{d\times K'}$ is the mapping matrix that transmit $\boldsymbol{h}_{t,n}$ into the embedding space of $\boldsymbol{z}_{t,n}$.

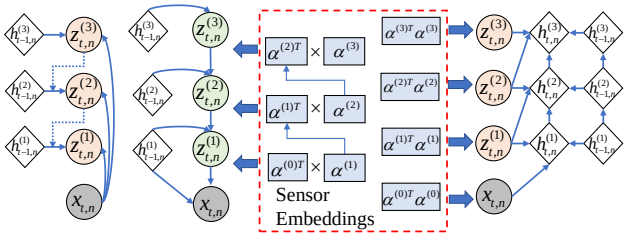
Compared with previous probabilistic generative network, EPN module discovers the latent semantic structure of each channel as an embedding vector, and provides a low-dimensional channel representation, thus capturing the relationships between each other according to the similarity of channel embeddings, and enhancing the robust representation learning in complex MTS scenarios.

3.3.2. Graph convolutional recurrent network

To consider both temporal dependencies and inter-relations, we introduce GCRN module by incorporating graph convolutional and recurrent structure into a combined framework. Specifically, with the latent representations and channel embeddings in Eq. (4), we first adopt data adaptive graph convolutional generation module (Bai et al., 2020) to infer the hidden inter-dependence of data automatically as:

$$\begin{aligned}
&\boldsymbol{A} = \text{Re}LU\left([\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}]^T [\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}]\right) \\
&\boldsymbol{H}_{t,n}^{(0)} = \ln([\boldsymbol{x}_{t,n}, \boldsymbol{z}_{t,n}]), \tilde{\boldsymbol{A}} = \boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{Q}^{-\frac{1}{2}} \\
&\boldsymbol{H}_{t,n}^{(1)} = \ln(1 + \exp(\boldsymbol{W}\tilde{\boldsymbol{A}}\boldsymbol{H}_{t,n}^{(0)}))
\end{aligned} \quad (5)$$

where $\boldsymbol{H}_{t,n}^{(0)}$ is the combination of input and probabilistic latent states. $[\cdot]$ means the concatenate operation. Similar as defining the graph by nodes similarity, the spatial dependencies between each pair of channels are inferred by multiplying their embeddings and referred to as a symmetric adjacent matrix $\boldsymbol{A}$. $\tilde{\boldsymbol{A}}$ is the normalized symmetric adjacent matrix with degree matrix Q. $\boldsymbol{W} \in \mathbb{R}^{(V+K)\times K'}$ is GCN filter. Generally speaking, we aggregate and propagate



(a) Inference (b) Generation (c) Stacked GCRN

Figure 2. Graphical illustration of each operation of the DVGCRN.

the structural and relational information of both the MTS and the latent states of probabilistic models into $\boldsymbol{H}_{t,n}^{(1)}$ via a GCN. To discover the complex temporal correlations, we further introduce a recurrent structure into our model and formalize as:

$$\boldsymbol{h}_{t,n} = f_{\boldsymbol{\theta}}(\boldsymbol{H}_{t,n}^{(1)}, \boldsymbol{h}_{t-1,n}) \tag{6}$$

where $f_{\boldsymbol{\theta}}(\cdot)$ is a non-linear transition function, which can be implemented by an LSTM with parameter $\boldsymbol{\theta}$. The latent states $\boldsymbol{h}_{t,n}^{(l)}$ are updated with $\boldsymbol{H}_{t,n}^{(1)}$ and $\boldsymbol{h}_{t-1,n}$, thus to incorporate both relational and temporal information.

3.4. Deep variational graph convolutional recurrent network

To further improve the generalization capacity, a straightforward extension of VGCRN is to apply hierarchical prior on the EPN module and stacked structure on the GCRN module. Therefore, as shown in Fig. 2, we construct a multi-layer probabilistic model named DVGCRN.

3.4.1. Deep Embedding-Guided Probabilistic Generative Network

To capture the hierarchical structural and relational characteristics, as shown in Fig. 2 (b), we formulate DEPN into a hierarchical generation process and incorporate multi-level embedding channels at each layer as:

$$\begin{aligned}
&\boldsymbol{W}_{h\mu}^{(l)} = \text{softmax}({\boldsymbol{\alpha}^{(l)}}^T \boldsymbol{\beta}^{(l)}) \\
&\boldsymbol{W}_{z\mu}^{(l)} = \text{softmax}({\boldsymbol{\alpha}^{(l-1)}}^T \boldsymbol{\alpha}^{(l)}) \\
&\boldsymbol{\mu}_{t,n}^{(L)} = f\left(\boldsymbol{W}_{h\mu}^{(L)} \boldsymbol{h}_{t-1,n}^{(L)}\right) \\
&\boldsymbol{\mu}_{t,n}^{(l)} = f\left(\boldsymbol{W}_{z\mu}^{(l+1)} \boldsymbol{z}_{t,n}^{(l+1)} + \boldsymbol{W}_{h\mu}^{(l)} \boldsymbol{h}_{t-1,n}^{(l)}\right), ... \\
&\boldsymbol{z}_{t,n}^{(l)} \sim \left(\boldsymbol{\mu}_{t,n}^{(l)}, \text{diag}\left(\boldsymbol{\sigma}_{t,n}^{(l)}\right)\right), l = 1, ..., L
\end{aligned} \tag{7}$$

The generation process from $\boldsymbol{z}_{t,n}^{(1)}$ to $\boldsymbol{x}_{t,n}$ is the same as Eq. 4. $\boldsymbol{z}_{t,n}^{(l)} \in \mathbb{R}^{K_l}$ denotes the Gaussian distributed probabilistic latent variables at layer l. $\boldsymbol{h}_{t,n}^{(l)} \in \mathbb{R}^{d_l}$ denotes the deterministic latent states of SGCRN at the l-th layer. We convert channels at different layers into a shared probabilistic embedding space as Eq. (1), (2) and (3), and define the factor loading matrix $\boldsymbol{W}_{z\mu}^{(l)} \in \mathbb{R}^{K_l \times K_{l-1}}$ at layer $l < L$ as the inner product of channel embeddings at the l-th and $(l-1)$-th layers, which can not only capture the relationships between channels at adjacent layers, but also help to inject information to higher layers as in Duan et al. (2021), thus ensuring the expressive ability of deep structure. $\boldsymbol{W}_{z\mu}^{(l)} \in \mathbb{R}^{K_l \times d_l}$, refers to a matrix that transits temporal information from $\boldsymbol{h}_{t,n}^{(l)}$ to the probabilistic latent variables. $\beta^{(l)}$ is the mapping matrix at layer l.

3.4.2. Stacked Graph Convolutional Recurrent Network

To consider multilevel temporal dependencies and interrelations, we extend GCRN into a multilayer network named SGCRN, as shown in Fig. 2(c). For the first layer of SGCRN, we formulate it in the same way as Eqs. (5) and (6). Then, the formulation of higher layers can be expressed as:

$$\begin{aligned}
&\boldsymbol{A}^{(l)} = \text{ReLU}\left([\boldsymbol{\alpha}^{(l-1)}, \boldsymbol{\alpha}^{(l)}]^T [\boldsymbol{\alpha}^{(l-1)}, \boldsymbol{\alpha}^{(l)}]\right) \\
&\tilde{\boldsymbol{H}}_{t,n}^{(l)} = \ln({\boldsymbol{z}_{t,n}^{(l-1)}}^T, {\boldsymbol{z}_{t,n}^{(l)}}^T), \tilde{\boldsymbol{A}}^{(l)} = \boldsymbol{Q}^{-\frac{1}{2}} \boldsymbol{A}^{(l)} \boldsymbol{Q}^{-\frac{1}{2}} \\
&\boldsymbol{H}_{t,n}^{(l)} = \ln(1 + \exp(\boldsymbol{W}^{(l)} \tilde{\boldsymbol{A}}^{(l)} \tilde{\boldsymbol{H}}_{t,n}^{(l)})), l = 2, ..., L \\
&\boldsymbol{h}_{t,n}^{(l)} = \text{LSTM}^{(l)}([\boldsymbol{H}_{t,n}^{(l)}, \boldsymbol{h}_{t,n}^{(l-1)}], \boldsymbol{h}_{t-1,n}^{(l)})
\end{aligned} \tag{8}$$

where $\boldsymbol{z}_{t,n}^{(l)}$ is the latent state at layer l, and we introduce a GCN at each layer to combine the structure and relation information within $\boldsymbol{z}_{t,n}^{(l)}$. To better utilize the dependencies between channels at both the same and adjacent layers, we combine $\boldsymbol{z}_{t,n}^{(l)}$ with $\boldsymbol{z}_{t,n}^{(l-1)}$ as the input of GCN at layer l to construct the adjacent matrix as $\boldsymbol{A}^{(l)}$. The $\boldsymbol{W}^{(l)} \in \mathbb{R}^{(K_{l-1}+K_l) \times K_l^{'}}$ refers to the GCN's filter at layer l. After getting $\boldsymbol{H}_{t,n}^{(l)}$ with GCN at each layer, a stacked LSTM is introduced to further characterize multi-level and long-range temporal dependencies. Besides, to obtain better time-series representations, we assign SGCRN prediction task as:

$$\hat{\boldsymbol{x}}_{T,n} = f(\boldsymbol{W}_p \boldsymbol{s}_{T-1,n}), \boldsymbol{s}_{T-1,n} = [\boldsymbol{h}_{T-1,n}^{(1)}, ..., \boldsymbol{h}_{T-1,n}^{(L)}] \tag{9}$$

where $\boldsymbol{s}_{T-1,n} \in \mathbb{R}^{d_1+...+d_L}$ is the concatenation of latent states across all layers and $\boldsymbol{W}_p \in \mathbb{R}^{V \times (d_1+...+d_L)}$ denotes a weight matrix for predicting $\boldsymbol{x}_{T,n}$.

In summary, DVGCRN integrates DEPN to enhance generalization capacity with hierarchical prior and capture the dependencies between channels in embedding space, and the SGCRN to characterize multilevel relational and temporal dependencies into a well-defined hybrid Bayesian framework with the shared probabilistic channel embeddings.

3.4.3. Upward-Downward Inference Scheme

Following VAE-based models (Kingma & Welling, 2014; Rezende et al., 2014), a flexible variational distribution $q(\boldsymbol{z}_{t,n}|\boldsymbol{x}_{t,n})$ could be defined to approximate the true posterior distribution $p(\boldsymbol{z}_{t,n}|-)$. To avoid collapsing of the stochastic latent variables at higher layers into their prior caused by the hierarchical structure (Zhou et al., 2015; Guo

et al., 2020), we not only construct a deterministic-upward path that links the recurrent latent states to the multi-layer latent representations from $l = 1$ to $l = L$, but also parameterize a mapping from input $\boldsymbol{x}_{t,n}$ to them:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{t,n}^{(l)} &= f(\boldsymbol{C}_{x\mu}^{(l)}\boldsymbol{x}_{t,n} + \boldsymbol{C}_{h\mu}^{(l)}\boldsymbol{h}_{t-1,n}) \\ \hat{\boldsymbol{\sigma}}_{t,n}^{(l)} &= f(\boldsymbol{C}_{x\sigma}^{(l)}\boldsymbol{x}_{t,n} + \boldsymbol{C}_{h,n}^{(l)}\boldsymbol{h}_{t-1,n})\end{aligned} \tag{10}$$

Where $\{\boldsymbol{C}_{x\mu}^{(l)}, \boldsymbol{C}_{x\sigma}^{(l)}\} \in \mathbb{R}^{V\times d_l}$ and $\{\boldsymbol{C}_{h\mu}^{(l)}, \boldsymbol{C}_{h\sigma}^{(l)}\} \in \mathbb{R}^{d_l^{'}\times d_l}$ are all learnable parameters of the inference network. Then, DVGCRN combines the obtained latent features with the prior from the stochastic-downward path to construct the variational posterior of latent states from $l = L$ to $l = 1$:

$$\begin{aligned}q\left(\boldsymbol{z}_{t,n}^{(l)}\right) &= \mathcal{N}(\boldsymbol{\mu}_{t,n}^{(l)}, \text{diag}\left(\boldsymbol{\sigma}_{t,n}^{(l)}\right)) \\ \boldsymbol{\mu}_{t,n}^{(l)} &= \text{linear}(\hat{\boldsymbol{\mu}}_{t,n}^{(l)} + \boldsymbol{W}_{z\mu}^{(l)}\boldsymbol{z}_{t,n}^{(l+1)}) \\ \boldsymbol{\sigma}_{t,n}^{(l)} &= \text{Softplus}(\text{linear}(\hat{\boldsymbol{\sigma}}_{t,n}^{(l)} + 1))\end{aligned} \tag{11}$$

Based on the structure of the inference network, the posterior of probabilistic latent variables of DVGCRN are approximated by combing the bottom-up likelihood and up-bottom prior information from the generative distribution, thus to enable richer latent representations for DVGCRN.

Moreover, we conclude the properties of DCGCRN in Section. F in the Appendix, and all properties ensure more expressive and robust representations of normal MTS patterns, thus achieves accurate detection for both ground-truth and misjudgment regions in Fig. 1.

3.5. Model training

We combine prediction and reconstruction tasks into the optimization objective. Specifically, the forecasting-based model focuses on single-time step prediction, while the reconstruction-based model learns a latent representation for entire MTS. For DVGCRN, given channel embedding $\boldsymbol{\alpha}^{(1:L)}$ and model parameters referred to as $\boldsymbol{W}^{(1;L)}$, the marginal likelihood of the MTS dataset $\mathcal{D}$ is defined as:

$$\begin{aligned}P(\mathcal{D}|\{\boldsymbol{\alpha}^{(l)}, \boldsymbol{W}^{(l)}\}_{l=1}^{L}) &= \int \prod_{t=1}^{T} p(\boldsymbol{x}_{t,n}|\boldsymbol{z}_{t,n}^{(1)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}) \\ &+ [\prod_{l=1}^{L}\prod_{t=1}^{T} p(\boldsymbol{z}_{t,n}^{(l)}|\boldsymbol{z}_{t,n}^{(l-1)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\alpha}^{(l-1)})] \\ &+ [\prod_{l=1}^{L} p(\boldsymbol{x}_{T,n}|\boldsymbol{h}_{1:T-1,n}^{(l)}, \boldsymbol{\alpha}^{(l)})] d\boldsymbol{z}_{1:T,n}^{1:L}\end{aligned}$$

where the first and the second terms are reconstruction loss, while the third one is prediction loss. The inference task is to learn the parameters of both the DEPN and DGCRN, as well as the shared channel embeddings. Similar to VAEs, the optimization objective of DVGCRN can be achieved by maximizing the evidence lower bound (ELBO) of the log marginal likelihood, which can be computed as:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N}\left[\sum_{t=1}^{T}\mathbb{E}_{q(\boldsymbol{z}_{t,n}^{(1)})}\left[\ln p\left(\boldsymbol{x}_{t,n}|\boldsymbol{z}_{t,n}^{(1)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}\right)\right]\right. \\ &+ \gamma\sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{z}_{t,n}^{(l)})}\left[\ln p\left(\boldsymbol{x}_{T,n}|h_{1:T-1,n}^{(l)}, \boldsymbol{\alpha}^{(l)}\right)\right] \\ &\left.- \sum_{t=1}^{T}\sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{z}_{t,n}^{(1)})}[\ln\frac{q\left(\boldsymbol{z}_{t,n}^{(l)}|\boldsymbol{x}_{t,n}, \boldsymbol{h}_{t-1,n}^{(l)}\right)}{p\left(\boldsymbol{z}_{t,n}^{(l)} \mid \boldsymbol{z}_{t,n}^{(l-1)}, \boldsymbol{\alpha}^{(l)}, \boldsymbol{\alpha}^{(l-1)}\right)}]\right]\end{aligned} \tag{12}$$

As we can see, different from traditional ELBO, which only consists of the expected log-likelihood of generative model ensuring reconstruction performance and the Killback–Leibler (KL) divergence that constrains the variational distribution $q(\boldsymbol{z}_{t,n}^{(l)})$ to be close to its prior $p(\boldsymbol{z}_{t,n}^{(l)})$, the ELBO for DVGCRN also incorporates the prediction loss by SGCRN to ensure the expressive time-series representation. $\gamma > 0$ is a hyper-parameter to balance the forecasting-based error and the reconstruction-based probability, which is chosen by grid search on the validation set. The parameters of DEPN module, inference module, SGCRN module and channel embeddings, defined as $\{\boldsymbol{\Psi}^{(l)}\}_{l=1}^{L}$, $\{\boldsymbol{\Omega}^{(l)}\}_{l=1}^{L}$, $\boldsymbol{\Theta}$ and $\{\boldsymbol{\alpha}^{(l)}\}_{l=1}^{L}$ respectively, in DVGCRN are jointly updated by stochastic gradient descent (SGD), as described in Algorithm. 1.

Algorithm 1 Upward-Downward Autoencoding Variational Inference for DVGCRN

Set mini-batch size as M, the number of convolutional filters K and hyperparameters;
Initialize the parameters of inference networks $\boldsymbol{\Omega}$, DEPN $\boldsymbol{\Psi}$, SGCRN $\boldsymbol{\theta}$ and the channel embeddings $\boldsymbol{\alpha}^{(0:L)}$;
repeat
  Randomly select a mini-batch of M MTS consist of T subsequences to form a subset $\{\boldsymbol{x}_{1:T,i}\}_{i=1}^{M}$;
  Draw random noise $\{\boldsymbol{\epsilon}_{t,n}^{(l)}\}_{t=1,n=1,l=1}^{T,M,L}$ from uniform distribution for sampling latent states $\{z_{t,n}^{(l)}\}_{t=1,n=1,l=1}^{T,M,L}$;
  Calculate $\nabla L\left(\boldsymbol{\Omega}, \Psi; X, \boldsymbol{\epsilon}_{t,i}^{(l)}, \boldsymbol{\theta}, \boldsymbol{\alpha}^{(0:L)}\right)$ according to Eq. (12), and update encoder parameters $\boldsymbol{\Omega}$ and decoder parameters $\boldsymbol{\Psi}$ jointly;
until convergence
return global parameters $\{\boldsymbol{\Omega}, \boldsymbol{\Psi}, \boldsymbol{\theta}, \boldsymbol{\alpha}^{(0:L)}\}$.

4. Anomaly detection based on DVGCRN

4.1. Overview of the framework

The overview framework for MTS unsupervised detection employing DVGCRN is shown in Fig. 3. It contains three components. The first one pre-processes the original multivariate time series data so that it can be used to train the model. Specifically, normalization and sliding time window approaches (Dai et al., 2021) are adopted. We then employ DVGCRN to learn the representations of MTS. Finally, anomalies are detected in terms of the reconstruction and

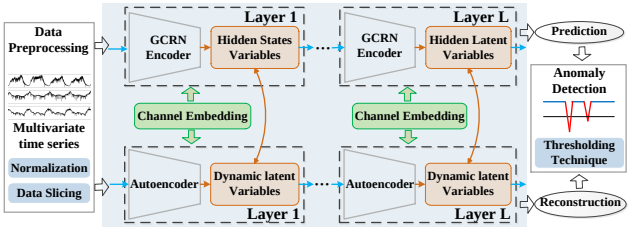


Figure 3. Overall framework of anomaly detection for MTS based on DVGCRN.

prediction probability inferred from the DVGCRN.

4.2. Anomaly score

Since the model is trained to learn normal patterns of multivariate time series, the more an observation follows normal patterns, the more likely it can be reconstructed and predicted well with higher confidence. Hence, we apply the reconstruction probability and prediction error of x_t as the anomaly score to determine whether an observed variable is anomalous or not (An & Cho, 2015), and it is computed as:

$$\begin{aligned} \mathcal{S}_{t,n} &= \left(\mathcal{S}^r_{t,n} + \gamma(-\mathcal{S}^p_{t,n})\right)/(1+\gamma) \\ \mathcal{S}^r_{t,n} &= \log p(\boldsymbol{x}_{t,n}|\boldsymbol{z}_{t,n}), \mathcal{S}^p_{t,n} = (\boldsymbol{x}_{t,n} - \hat{\boldsymbol{x}}_{t,n})^2 \end{aligned} \tag{13}$$

where $\mathcal{S}^r_{t,n}$ and $\mathcal{S}^p_{t,n}$ are reconstruction and prediction score, respectively. An observation x_t will be classified as anomalous if $\mathcal{S}_{t,n}$ is below a specific threshold. To better utilize the multi-layer representations and improve the robustness, we are motivated to modify the reconstruction anomaly score as an united conditional probability as:

$$\begin{aligned} \hat{\mathcal{S}}^r_{t,n} &= \frac{1}{L}\log p\left(\boldsymbol{x}_{t,n}, \boldsymbol{z}^{(1)}_{t,n}, ..., \boldsymbol{z}^{(L-1)}_{t,n} \mid \boldsymbol{z}^{(L)}_{t,n}\right) \\ &= \frac{1}{L}(\log p(\boldsymbol{x}_{t,n}|\boldsymbol{z}^{(1)}_{t,n}) + \sum_{l=1}^{L-1}\log(\boldsymbol{z}^{(l)}_{t,n} \mid \boldsymbol{z}^{(l+1)}_{t,n})) \end{aligned} \tag{14}$$

From a practical point of view, we use the Peaks-Over-Threshold (POT) (Siffer et al., 2017) approach to help select threshold. In our case, the lower anomaly scores are more likely considered to be extreme values since the lower anomaly score, the greater chance it belongs to outlier.

5. Experiment

5.1. Experiment setting

Baseline methods: We compare DVGCRN with state-of-the-art online anomaly detection methods for MTS: *1)* Non-dynamical methods, including VAE (Kingma & Welling, 2014), a basic probabilistic generative model; Ladder VAE (Sønderby et al., 2016), a deep VAE structured model; *2)* Dynamical methods, including LSTM-NTD (Hundman et al., 2018), a deterministic recurrent autoencoding model; VRNN (Chung et al., 2015), a probabilistic dynamic model for describing high-dimensional sequences; OmniAnomaly (Su et al., 2019), a stochastic RNN-based model; SDFVAE (Dai et al., 2021), a static and dynamic factorized VAE-based framework. *3)* relational methods, including GNNs (Deng & Hooi, 2021), a graph networks for anomaly detection; InterFusion (Li et al., 2021) that considers inter-relationships with convolutional structure.

Hyper-parameters: For baselines, we use the official parameter settings reported in their experiments. For our DVGCRN, we implement it using Pytorch library with mini-batch 256 for 120 running epochs. The Adam optimizer is used with learning rate 0.0002. Besides, we set temperature coefficient β in ELBO equation increasing from 0 to 1 during the first 10 training epochs. And the probability p associated with POT is set to be 0.004 empirically.

Datasets: We conduct extensive experiments on twofold datasets: *1)* the DND, a real-world multivariate KPI dataset; *2)* the SMD, the MSL, and the SMAP, three public datasets released by (Su et al., 2019) and (Hundman et al., 2018), details of which can be found in the Appendix.

Hardware platform: All experiments are performed on workstation equipped with a CPU i7-10700 and accelerated by one GPU NVIDIA RTX 3090 with 24GB VRAM.

Evaluation metrics: Similar to previous studies (Dai et al., 2021; Su et al., 2021; 2019), we employ Precision, Recall, and F1-score as evaluation metrics to compare the performance of the different methods. Particularly, F1 is deemed as a comprehensive indicator since it balances precision and recall. And our code is available at https://github.com/BoChenGroup/DVGCRN.

5.2. Quantitative comparison

The influence of model parameters: In this part, we discuss the effect of model parameters on anomaly detection performance. Setting the network structure as $[15, 10, 5]$, we first evaluate the effect of window size T, the data preprocessing parameter introduced in Sec. 4.1, which determines the range of temporal dependencies. Fig. 4 (left) shows the variation of detection accuracy of our model with T ranging from 5 to 25. Clearly, as the length of T increases, there is a common trend for the detection accuracy of DVGCRN with different layers. Specifically, it grows first and then keeps stabilized, which can be attributed to the limited capacity of models in the length of temporal dependencies they can capture. Particularly, for the sake that DVGCRN with deeper layer can get longer range of temporal dependencies, thus achieving more performance improvement with the increase of T. Besides, fixing $T = 20$, we investigate the effect of network size, including L and $\{d_l\}_{l=1}^L$, on the performance of anomaly detection and report the results in Fig. 4 (middle). As we can see, there is a clear trend of improvement in anomaly detection accuracy by increasing the network depth given a limited first-layer width, indicating the effectiveness of hierarchical structure in enhancing the representation and generalization power. However, with the increase of the hidden-layer width given

Table 1. Detection performance of different methods. The best results are bolded.

Methods	Network Size	Latent Size	DND			SMD			MSL			SMAP		
			P	R	F1-score	P	R	F1-score	P	R	F1-score	P	R	F1-score
VAE (Kingma & Welling, 2014)	-	15	0.913	0.635	0.749	0.989	0.685	0.809	0.781	0.813	0.797	0.755	0.737	0.746
ladder VAE (Sønderby et al., 2016)	-	15-10-5	0.813	0.772	0.765	0.990	0.753	0.855	0.842	0.823	0.832	0.792	0.775	0.783
LSTM-NTD (Hundman et al., 2018)	20	-	0.585	0.614	0.599	0.568	0.644	0.604	0.593	0.537	0.564	0.896	0.885	0.891
VRNN (Chung et al., 2015)	20	15	0.802	0.777	0.789	0.970	0.795	0.874	0.884	0.902	0.893	0.805	0.821	0.813
GNN (Deng & Hooi, 2021)	20	-	0.826	0.796	0.811	0.829	0.964	0.891	0.881	0.889	0.885	0.812	0.944	0.873
OmniAnomaly (Su et al., 2019)	20	15	0.919	0.723	0.809	0.819	0.968	0.887	0.887	0.912	0.899	0.742	0.978	0.843
InterFusion (Li et al., 2021)	-	15	0.853	0.782	0.816	0.829	0.968	0.893	0.882	0.926	0.903	0.889	0.910	0.899
SDFVAE (Dai et al., 2021)	20	15	0.964	0.711	0.818	0.882	0.926	0.903	0.853	0.894	0.873	0.884	0.908	0.896
VGCRN-rec	20	15	0.902	0.751	0.820	0.959	0.835	0.893	0.881	0.924	0.902	0.890	0.911	0.901
VGCRN	20	15	0.822	0.831	0.826	0.959	0.850	0.901	0.889	0.920	0.904	0.886	0.930	0.908
DVGCRN-layer2	20-15	15-10	0.929	0.754	0.832	0.960	0.850	0.902	0.890	0.922	0.906	0.905	0.915	0.910
DVGCRN-layer3	20-15-10	15-10-5	0.930	0.756	0.834	0.950	0.876	0.912	0.886	0.930	0.908	0.908	0.916	0.912
DVGCRN-layer3-M	20-15-10	15-10-5	0.929	0.782	**0.849**	0.950	0.883	**0.915**	0.888	0.941	**0.914**	0.916	0.920	**0.914**

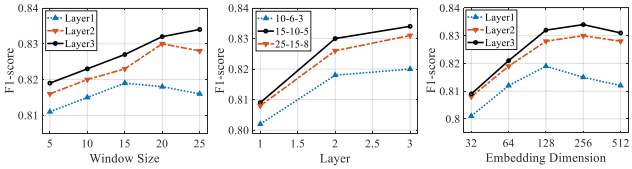


Figure 4. F1-score of the DVGCRN on the DND as a function of window size (left), network size (middle) and embedding dimension (right).

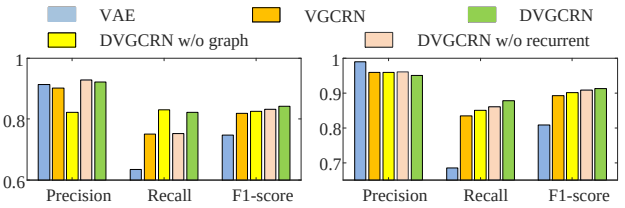


Figure 5. Ablation studies of the DVGCRN on the DND (left) and the SMD (right) datasets.

a fixed depth, the performance for anomaly detection of DVGCRN first increases and then decreases, which can be attributed to that the hidden states with the higher dimension contain more useful information and thus enable DEPN to learn more robust representation of MTS, while higher latent dimension may results in higher computational cost, making the model harder to optimize and easier to over-fit. In addition, Fig.4 (right) also shows the effect of different embedding dimension d on DVGCRN, which not only influences the quality of the learned graph but also decides the parameter diversity in DEPN. DVGCRN with 2 or 3 layers achieves the best performance when the embedding dimension is 256, while single-layer DVGCRN performs best when $d = 128$, indicating that the excessively small and large channel embedding dimension will also lead to poorer performance for the same reason with network size discussed above.

Anomaly detection performance: We compare the proposed models with baselines introduced in Sec. 5.1 F1-score, and report the average results on five independent runs in Table.1. It is obvious that probabilistic methods achieve better results than deterministic ones with the same architecture since they consider the stochastic within MTS. Both recurrent and graph structures can boost the performance, indicating that the effectiveness of temporal and relational characteristics on learning normal patterns of MTS. VGCRN achieves better results than other probabilistic methods for incorporating the two structures. Besides, VGCRN-rec refers to VGCRN optimized with only reconstruction loss, it underperforms the VGCRN, illustrating that the effectiveness of our loss in Eq. (12). The details please refer to the Appendix. DVGCRN with three layers achieves the best detection performance among all methods, which demonstrates that the richer structure and relation information provided by the deep probabilistic models can further improve the learning of robust representations of MTS. "-M" denotes using the developed anomaly score in Eq.14, which does help according to the results. Moreover, we also list the F1-best performance of different methods in the Appendix.

Ablation study: We conduct ablation study to analyze the roles of graph, recurrent and hierarchical structure in the proposed models by comparing the performance of DVGCRN, VGCRN, DVGCRN w/o graph by removing $\boldsymbol{\alpha}_{(l)}$, DVGCRN w/o recurrent by setting T = 1 and the basic VAE. Observations can be drawn from Fig. 5 that each component incorporated into DVGCRN can bring significant improvement on performance.

5.3. Qualitative analysis

Anomaly score: To show the efficiency of different models in capturing the normal patterns of time series intuitively, we visualize the anomaly score of the case study. Firstly, we compare anomaly scores between VGCRN, DVGCRN with three layers, and some baselines, including LSTM-NTD, GNN, OmniAnomaly, and InterFusion. The results are visualized in Fig. 7. As deterministic methods, LSTM-NTD and GNN get more turbulent anomaly scores since they ignore the stochastic of MTS. Considering the inter-relationship within MTS, GNN exhibits more distinct spikes in the anomaly regions than LSTM-NTD. For probabilistic methods, considering both temporal and inter-metric dependencies, InterFusion outperforms OmniAnomaly. Modeling relational and temporal characteristic in the generative process and considering their variations, VGCRN realizes

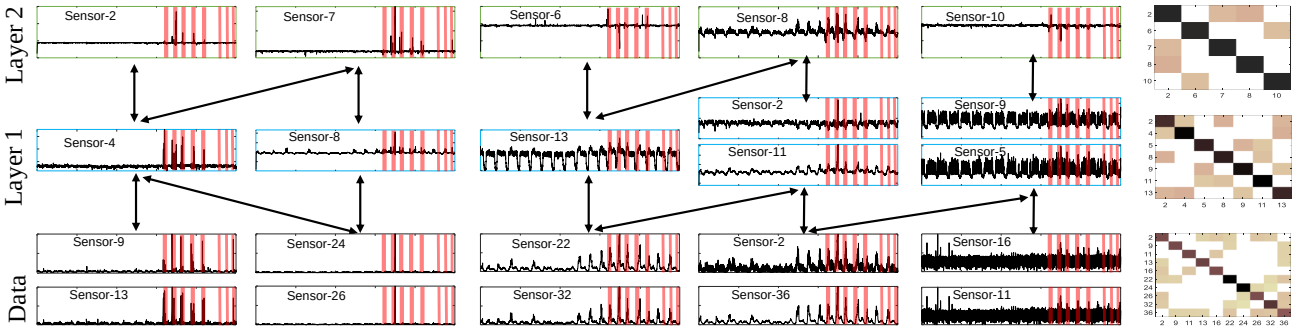


Figure 6. Visualizations for latent representations of data and the stochastic latent representations (y and x axis indicate amplitude and time) and channel relationships at different layers of the DVGCRN on the SMD mechine-1-1 dataset.

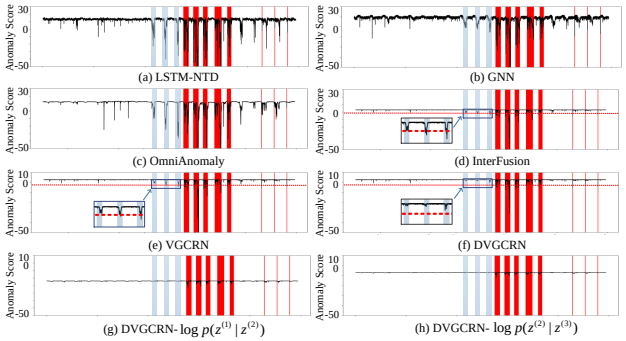


Figure 7. Case studies of reconstruction anomaly scores on the SMD mechine-1-1 dataset. Regions highlighted in red represent ground-truth anomaly segments.

smoother anomaly scores than previous methods. Moreover, DVGCRN has better representation power with the help of hierarchical architecture, leading to more stable anomaly scores when compared with its shallow peers, such as the regions highlighted in blue in Fig. 7, while it exhibits considerable spikes in anomaly regions. Anomaly scores further demonstrate the capability of DVGCRN in learning normal patterns of complex MTS, which echoes the numerical results in Table 1. Moreover, we visualize log-likelihoods of latent states in Fig. 7 (g) and (h). It is interesting to notice that they are stable at normal time steps and turbulent at abnormal ones, which tends to result in the stationary of modified anomaly scores. This phenomena illustrates the superiority of the multi-layer structure and the developed anomaly score in Eq. 14.

Latent representations and relationships: To better demonstrate the effectiveness of hierarchical and graph structures, we further visualize the learned latent representations and corresponding relational matrices of different layers, which are calculated by the inner product of channel embedding vectors, in Fig. 6. The latent representations at low layer are specific and vary intensely, while they vary smoothly as the depth increases, which is critical for DVGCRN to capture the temporal interactions between multiple closely time steps. Besides, there are characteristic fluctuations in anomaly time steps from bottom to top layers, indicating the effectiveness of features at each layer. Moreover, Fig. 6 (right) presents a subset of the corresponding relational matrices to these channels at each layer. As we can see, with the increase of the layers, the relational matrices exhibit stronger diagonal property, meaning that the feature channels are likely to be dependent on themselves, which matches the characteristic of DVGCRN that channels at higher layers have the ability to cover long-time dependencies and contain more general information. Moreover, as DVGCRN mapping channels at different layers into the shared embedding space, it can also capture the relationships among channels at different layers by $\text{softmax}(\boldsymbol{\alpha}^{(l-1)T}\boldsymbol{\alpha}^{(l)})$, which is denoted with the connecting line in Fig. 6. Thus injecting the learned knowledge of a lower layer to a higher layer, which can help the higher layer optimization and derive better representations for the normal patterns of MTS.

6. Conclusion

We propose VGCRN as a new unsupervised probabilistic method for anomaly detection of MTS, which combines our developed novel EPN with GCRN into a unified framework by a sharing embedding channel, thus learning the robust representations of MTS by considering both temporal, interrelationship and stochasticity characteristics. Then, we extend VGCRN to a deep version, DVGCRN, which is able to explore the hierarchical information of MTS. Besides, for efficient and accurate inference, we propose an upward-downward inference scheme combined with a hybrid reconstruction and prediction optimization target. Through both qualitative and quantitative experiments on real-world and public datasets, our models are shown to achieve promising results on anomaly detection and diagnosis.

Acknowledgments

Bo Chen acknowledges the support of NSFC (U21B2006 and 61771361), Shaanxi Youth Innovation Team Project, the 111 Project (No. B18039) and the Program for Oversea Talent by Chinese Central Government. Mingyuan Zhou acknowledges the support of U.S. National Science Foundation under Grant IIS-1812699.

References

An, J. and Cho, S. Variational autoencoder based anomaly detection using reconstruction probability. *Technical Report. SNU Data Mining Center*, pp. 1–8, 2015.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. USAD: unsupervised anomaly detection on multivariate time series. In *ACM SIGKDD '20*, pp. 3395–3404, 2020.

Bai, L., Yao, L., Kanhere, S. S., Wang, X., and Sheng, Q. Z. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. In *IJCAI*, pp. 1981–1987, 2019.

Bai, L., Yao, L., Li, C., Wang, X., and Wang, C. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, 2020.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009.

Chen, W., Xu, H., Li, Z., Pei, D., Chen, J., Qiao, H., Feng, Y., and Wang, Z. Unsupervised anomaly detection for intricate kpis via adversarial training of VAE. In *IEEE INFOCOM*, pp. 1891–1899, 2019.

Chen, W., Wang, C., Chen, B., Liu, Y., Zhang, H., and Zhou, M. Bidirectional convolutional poisson gamma dynamical systems. In *NeurIPS*, 2020.

Chen, Y., Mahajan, R., Sridharan, B., and Zhang, Z. A provider-side view of web search response time. In *ACM SIGCOMM '13*, 2013.

Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. In *ICLR*, 2021.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. In *NeurIPS*, pp. 2980–2988, 2015.

Dai, L., Lin, T., Liu, C., Jiang, B., Liu, Y., Xu, Z., and Zhang, Z. SDFVAE: static and dynamic factorized VAE for anomaly detection of multivariate CDN kpis. In *WWW*, pp. 3076–3086, 2021.

Deng, A. and Hooi, B. Graph neural network-based anomaly detection in multivariate time series. In *AAAI*, pp. 4027–4035, 2021.

Djurdjanovic, D., Lee, J., and Ni, J. Watchdog agent an infotronics-based prognostics approach for product performance degradation assessment and prediction. *Adv. Eng. Informatics*, 17:109–125, 2003.

Duan, Z., Wang, D., Chen, B., Wang, C., Chen, W., Li, Y., Ren, J., and Zhou, M. Sawtooth factorial topic embeddings guided gamma belief network. In *ICML*, volume 139, pp. 2903–2913, 2021.

Guo, D., Chen, B., Zhang, H., and Zhou, M. Deep poisson gamma dynamical systems. *in NeurIPS*, pp. 8451–8461, 2018.

Guo, D., Chen, B., Chen, W., Wang, C., Liu, H., and Zhou, M. Variational temporal deep generative model for radar HRRP target recognition. *IEEE Trans. Signal Process.*, 68:5795–5809, 2020.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Söderström, T. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *KDD*, pp. 387–395, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *ICLR*, 2014.

Len, R. A., Vittal, V., and Manimaran, G. Application of sensor network for secure electric energy infrastructure. *IEEE Transactions on Power Delivery*, 22:1021–1028, 2007.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., and Ng, S. MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks. In *Artificial Neural Networks and Machine Learning - ICANN*, pp. 703–716, 2019.

Li, Z., Zhao, Y., Han, J., Su, Y., Jiao, R., Wen, X., and Pei, D. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *KDD*, pp. 3220–3230, 2021.

Liu, D., Zhao, Y., Xu, H., Sun, Y., Pei, D., Luo, J., Jing, X., and Feng, M. Opprentice: Towards practical and automatic anomaly detection through machine learning. In *ACM IMC '15*, pp. 211–224, 2015.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., and Shroff, G. Lstm-based encoder-decoder for multi-sensor anomaly detection. In *ICML 2016 Anomaly Detection Workshop*, 2016.

Pang, G., Shen, C., Cao, L., and van den Hengel, A. Deep learning for anomaly detection: A review, 2020.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic back-propagation and variational inference in deep latent gaussian models. *ICML*, 2014.

Sepp, H. and Jurgen, S. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

Shon, T. and Moon, J. A hybrid machine learning approach to network anomaly detection. *Inf. Sci.*, 177(18):3799–3821, 2007.

Siffer, A., Fouque, P., Termier, A., and Largouet, C. Anomaly detection in streams with extreme value theory. In *ACM SIGKDD '17*, pp. 1067–1075, 2017.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *NeurIPS*, pp. 3738–3746, 2016.

Song, C., Lin, Y., Guo, S., and Wan, H. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *AAAI*, pp. 914–921, 2020.

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., and Pei, D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *SIGKDD*, pp. 2828–2837, 2019.

Su, Y., Zhao, Y., Sun, M., Zhang, S., and Pei, B. Detecting outlier machine instances through gaussian mixture variational autoencoder with one dimensional cnn. *IEEE Transactions on Computers*, PP(99):1–1, 2021.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., Chen, J., Wang, Z., and Qiao, H. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *WWW '18*, pp. 187–196, 2018.

Yamada, M., Kimura, A., Naya, F., and Sawada, H. Change-point detection with feature selection in high-dimensional time-series data. In *IJCAI '13*, pp. 1827–1833, 2013.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., and Chawla, N. V. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *AAAI*, pp. 1409–1416, 2019a.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., and Chawla, N. V. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *AAAI '19*, pp. 1409–1416, 2019b.

Zhang, H., Chen, B., Guo, D., and Zhou, M. Whai: Weibull hybrid autoencoding inference for deep topic modeling. *in ICLR*, 2018.

Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q. Multivariate time-series anomaly detection via graph attention network. In *ICDM*, pp. 841–850, 2020.

Zhou, M., Cong, Y., and Chen, B. The poisson gamma belief network. In *NeurIPS*, pp. 3043–3051, 2015.

Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. *J. Mach. Learn. Res.*, 17:163:1–163:44, 2016.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR '18*, 2018.

Appendix for "Deep Variational Graph Convolutional Recurrent Network for Multivariate Time Series Anomaly Detection"

A. Datasets

We conduct extensive experiments on four categories of datasets: one real-world dataset named the DND, a multivariate KPI dataset, and three public datasets named the SMD, the MSL and the SMAP that were released by (Su et al., 2019) and (Hundman et al., 2018), respectively. The basic statistical information of datasets is reported in Table 2. The DND, multivariate KPIs dataset, is collected from a large internet company in China. It contains 12 websites monitored with 36 KPIs individually. These websites are different from each other in types of services, e.g., Video on Demand (VoD) or live streaming video, etc. Besides, for each website, KPIs span about one and a half months and are collected every 60 seconds. In our experiments, for each website, the first half of the KPIs are used for training, while the last half are used for testing. Note that ground-truth anomalies at test time of the DND have been confirmed by human operators. For the public datasets, the Server Machine Dataset (SMD) is a real-world public dataset (Su et al., 2019) that contains the data from 28 server machines that are monitored by 38 KPIs individually. In addition, for each server machine, KPIs span about five weeks. the Mars Science Laboratory (MSL) rover dataset is also a real-world public and expert-labeled dataset from NASA (Hundman et al., 2018) containing the data of 27 entities and each of them is monitored by 55 metrics. Note that the other Soil Moisture Active Passive (SMAP) satellite dataset is also released by NASA (Hundman et al., 2018), and both MSL and SMAP are collected from spacecraft where the first dimension is the value of the telemetry channel, while the rest dimensions are command information that encoded as 0 or 1.

B. Time efficiency of DVGCRN

Table 3 shows the time efficiency of the DVGCRN in term of its training and testing time on the hardware platform introduced before. It can be seen from Table 3 that the DVGCRN can perform anomaly detection for a sample within one-tenth second versus the data collecting interval of 60 seconds. Hence, the DVGCRN can be deployed in an offline training and online detection manners (Dai et al., 2021; Su et al., 2019).

Table 2. Basic statistics of datasets

Statistics	DND	SMD	MSL	SMAP
Dimensions	12*36	28*38	27*55	55*25
Granularity (sec)	60	60	60	60
Training set size	344,843	708,405	58,317	135,181
Testing set size	344,844	708,420	73,729	427,617
Anomaly ratio (%)	3.44	4.16	10.72	13.13

Table 3. Training and testing time of DVGCRN

Datasets	Training times per epoch (min)	Testing times per sample (sec)
DND	31.2	0.068
SMD	28.8	0.069
MSL	2.1	0.084
SMAP	3.2	0.047

C. Balance between prediction and reconstruction

As discussed in Sec. 3.5, we combine both the prediction and the reconstruction loss on the DVGCRN, and introduce a parameter γ to balance the effects of them. Besides, γ are also used in Eq. 13 to get the anomaly score. Here, we evaluate the influence of γ to anomaly detection and report the results in Fig. 9. As we can see, excessively small and large γ will lead to the poor performance, illustrating the effectiveness of both the reconstruction and the prediction losses. Besides, for DND dataset, which has more noisy time series, high weight for reconstruction loss is good for detection, since probabilistic module can be robustness to noises and fluctuates. For SMD dataset, which has less noisy time series but contains more complex temporal dependencies, high weight for prediction loss is beneficial for learning better representation of time series.

D. More Ablation studies

To better illustrate the efficiency of our hierarchical embedding guided generation process of DVGCRN, we further do more ablation studies with F1-score and complexity in Fig. 9. Specifically, Fig. 9 (a) lists the F1-score of DVGCRN, DVGCRN w/o recurrent and DVGCRN w/o graph with different layers, while Fig. 9 (b) lists the the number of parameters changes with different layers. From Fig. 9, we

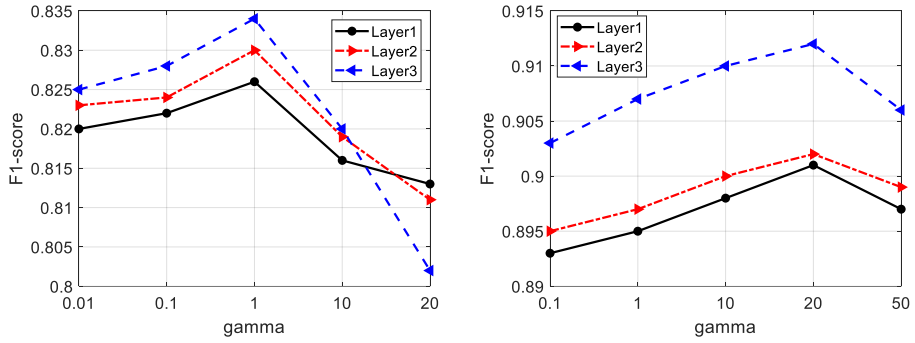


Figure 8. F1-scores of the DVGCRN on the DND (left) and the SMD (right) as a function of the balance parameter γ.

can conclude that: a) Affected by amount and complexity of the data, the performance of models will not always improve with the increasing number of layers; b) DVGCRN with one module but multiple layers are more complex than VGCRN, and they always underperform DVGCRN with multiple layers.

E. More anomaly detection performance

In Table 1, we use F1-score for performance evaluation and select threshold with POT approach, which is commonly used in OmniAnomaly, GNN and so on. Besides, we also test the F1-best performance, which is calculated by obtaining all F1-scores for each MTS by enumerating all thresholds and using the best F1-score as the final score, of different methods and list the results in Table 4. F1-best describes the best performance models can reach, but it is difficult to achieve in practice, while F1-score reflects models' generalization and robustness ability. As shown in Table 4, DVGCRN outperforms other methods on two metrics, illustrating the efficiency of the proposed method.

F. Model properties

Comparing with previous methods for modeling MTS, our proposed DVGCRN has the following properties:

***1)*:** Incorporating Gaussian-distributed channel embeddings into probabilistic generative process, modeling the stochasticity of different channels and their dependencies in MTS.
***2)*:** Enhancing generalization capacity with hierarchical structure and multi-layer channel embeddings, discovering multi-level representations and corresponding interdependencies of MTS;

***3)*:** Under the guidance of hierarchical statistic latent variables and channel embeddings by DEPN module, which exhibits different statistical properties across layers, and optimized with prediction loss, SGCRN can extract multi-scale temporal structures for better representation learning;

***4)*:** The prior of the latent variable in DEPN is conditioned on the previous latent states of SGCRN, thus to capture temporal dependencies for richer representations (Chung et al., 2015).

All of these properties ensure more expressive and robust representations of normal MTS patterns, thus achieves accurate detection for both ground-truth and misjudgment regions in Fig. 1.

Table 4. The detection performance of different methods.

Methods	DND		SMD		MSL		SMAP	
	F1-best	F1-score	F1-best	F1-score	F1-best	F1-score	F1-best	F1-score
GNN (2021 AAAI)	0.826	0.811	0.829	0.891	0.881	0.885	0.867	0.873
OmniAnomaly (2019 KDD)	0.816	0.809	0.954	0.887	0.903	0.899	0.859	0.843
SDFVAE (2021 WWW)	0.941	0.818	0.882	0.903	0.853	0.873	0.884	0.896
InterFusion (2021 KDD)	0.927	0.816	0.982	0.893	0.946	0.903	0.939	0.899
MTAD-GAT (2020 ICDM)	0.919	0.814	0.980	0.894	0.942	0.901	0.940	0.908
VGCRN (Ours)	0.940	0.826	0.986	0.901	0.950	0.904	0.941	0.908
DVGCRN (Ours)	**0.953**	**0.849**	**0.989**	**0.915**	**0.955**	**0.914**	**0.949**	**0.914**

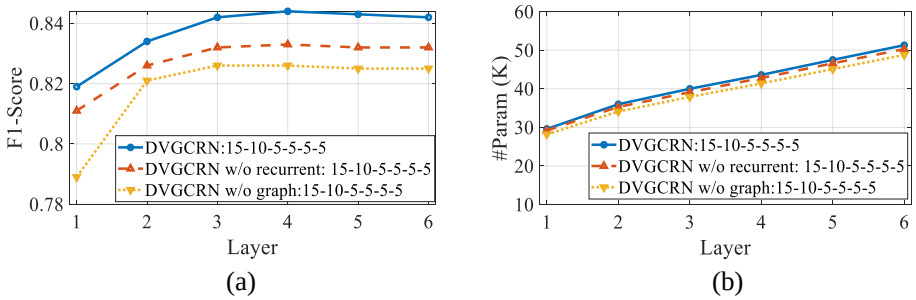


Figure 9. Ablation studies on the DND with the F1-score and #Param as a function of layers.