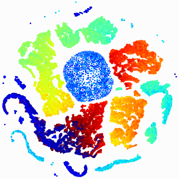
17th IEEE International Conference on Machine Learning and Applications



Unsupervised Anomaly Detection in Energy Time Series Data using Variational Recurrent Autoencoders with Attention

João Pereira & Margarida Silveira

Signal and Image Processing Group
Institute for Systems and Robotics
Instituto Superior Técnico (Lisbon, Portugal)

TÉCNICO LISBOA

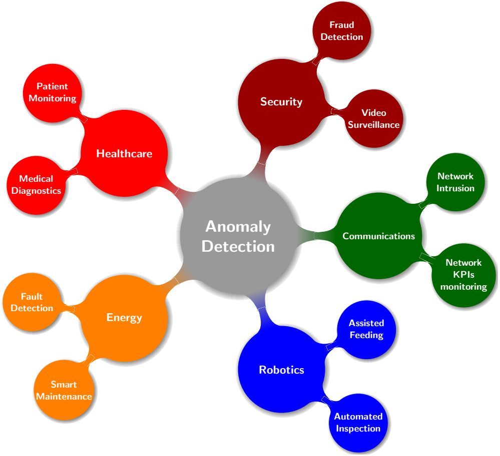
Institute for Systems and Robotics Lisboa

Orlando, Florida, USA

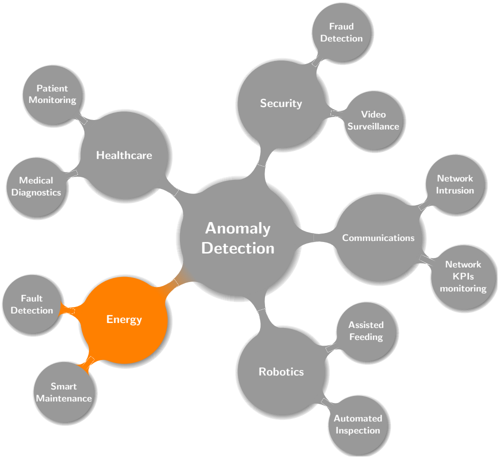
December 19th, 2018

Introduction

Anomaly detection is about finding patterns in data that do not conform to *expected* or *normal* behaviour.



Introduction



Main Challenges

- Most data in the world are **unlabelled**

anomaly labels

$$\text{Dataset } \mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)} \right) \right\}_{i=1}^{N}$$

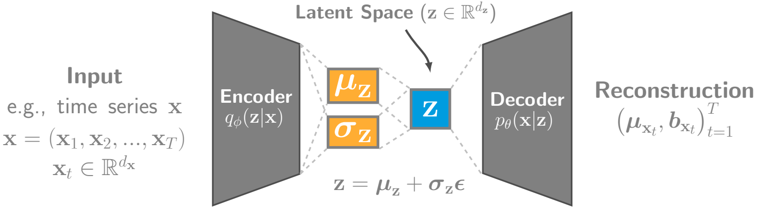
- Annotating large datasets is difficult, time-consuming and expensive

- Time series have temporal structure/dependencies

$$\mathbf{x} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \right) \quad , \qquad \mathbf{x}_t \in \mathbb{R}^{d_\mathbf{x}}$$

The Principle in a Nutshell

- Based on a **Variational Autoencoder**[1,2]

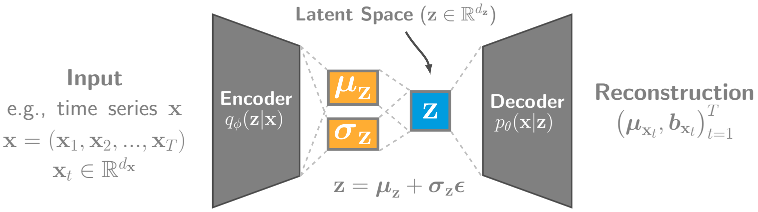


[1] Kingma & Welling, **Auto-Encoding Variational Bayes**, ICLR'14

[2] Rezende *et al.*, **Stochastic Backpropagation and Approximate Inference in Deep Generative Models**, ICML'14

The Principle in a Nutshell

- Based on a **Variational Autoencoder**[1,2]



- Train a VAE on data with mostly **normal** patterns;
- It reconstructs well **normal** data, while it fails to reconstruct **anomalous** data;
- The **quality of the reconstructions** is used as anomaly score.

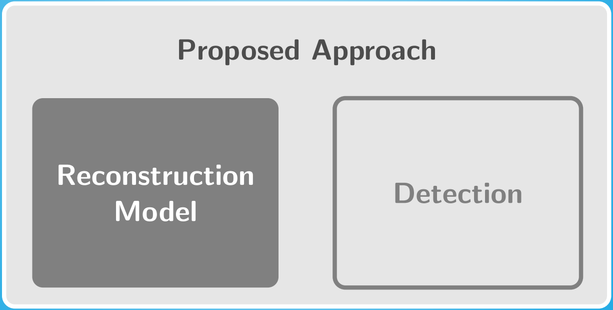
[1] Kingma & Welling, **Auto-Encoding Variational Bayes**, ICLR'14

[2] Rezende *et al.*, **Stochastic Backpropagation and Approximate Inference in Deep Generative Models**, ICML'14

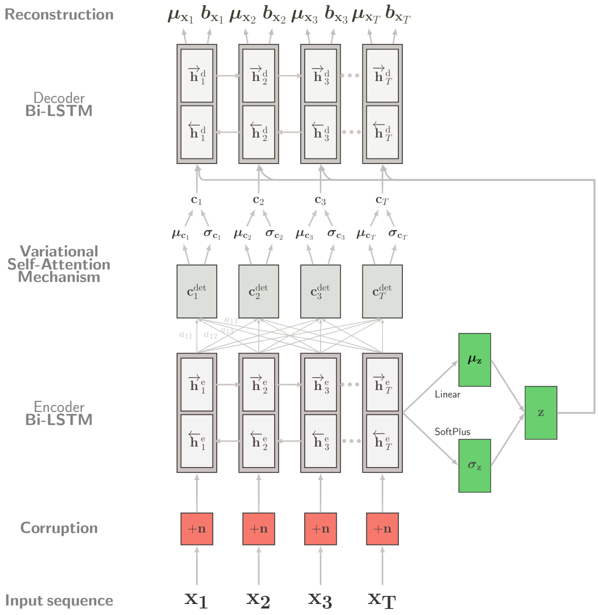
Proposed Approach

Reconstruction Model

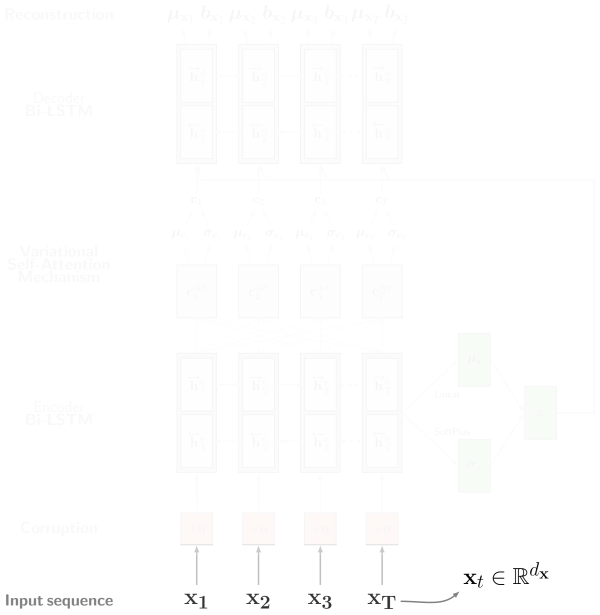
Detection


Proposed Approach
Reconstruction Model
Detection

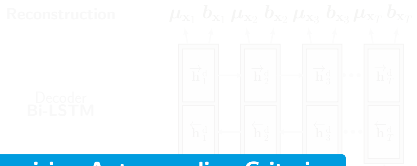
Reconstruction Model



Reconstruction Model



Reconstruction Model



Denoising Autoencoding Criterion

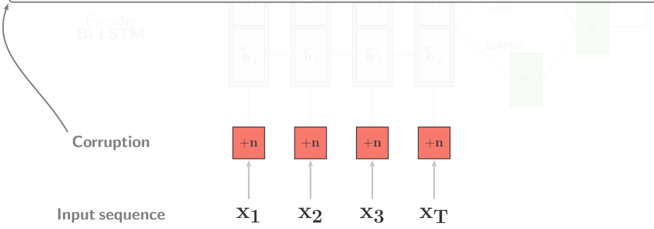
Corruption process: additive Gaussian noise

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \mathbf{x} + \mathbf{n} \quad , \quad \mathbf{n} \sim \mathrm{Normal}(\mathbf{0}, \sigma_{\mathbf{n}}^2 \mathbf{I})$$

♣

Vincent *et al.*, **Extracting and Composing Robust Features with Denoising Autoencoders**, ICML'08

Bengio *et al.*, **Denoising Criterion for Variational Auto-Encoding Framework**, ICLR'15



Learning temporal dependencies

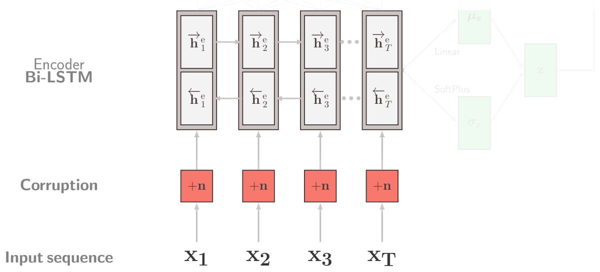
Bidirectional Long-Short Term Memory network

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right]$$

- 256 units, 128 in each direction
- Sparse regularization, $\Omega(\mathbf{z}) = \lambda \sum_{i=1}^{d_\mathbf{z}} |z_i|$

Hochreiter *et al.*, **Long-Short Term Memory**, Neural Computation'97

Graves *et al.*, **Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition**, ICANN'05



Reconstruction Model

Variational Latent Space

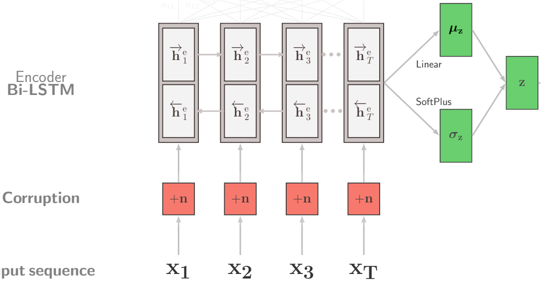
Variational parameters derived using neural networks

$$(\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\sigma}_{\mathbf{z}}) = \text{Encoder}(\mathbf{x})$$

Sample from the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$

$$\mathbf{z} = \boldsymbol{\mu}_{\mathbf{z}} + \boldsymbol{\sigma}_{\mathbf{z}} \odot \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$$

Kingma & Welling, **Auto-Encoding Variational Bayes**, ICLR'14



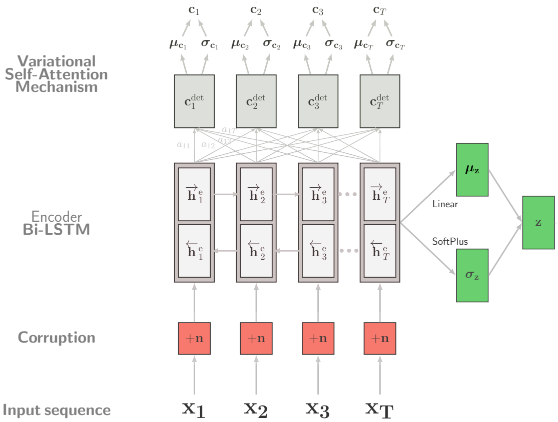
Reconstruction Model

Variational Self-Attention

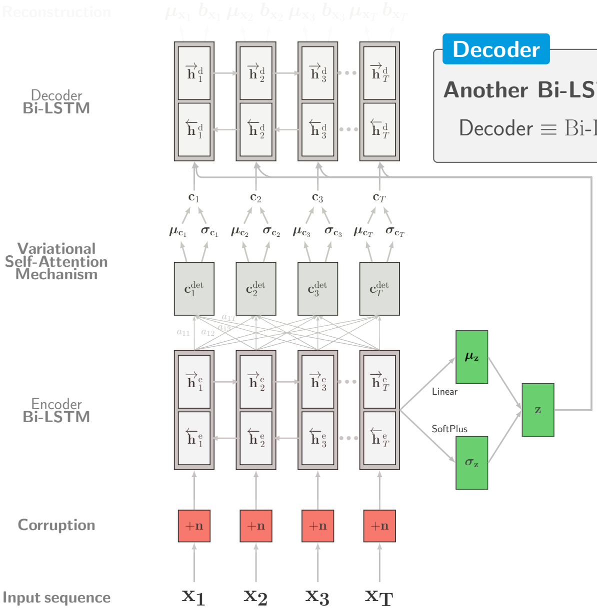
Combines self-attention with variational inference.

$$\mathbf{c}_t^{\text{det}} = \sum_{j=1}^{T} a_{tj}\mathbf{h}_j \qquad (\boldsymbol{\mu}_{\mathbf{c}_t}, \boldsymbol{\sigma}_{\mathbf{c}_t}) = \text{NN}(\mathbf{c}_t^{\text{det}}), \quad \mathbf{c}_t \sim \text{Normal}(\boldsymbol{\mu}_{\mathbf{c}_t}, \boldsymbol{\sigma}_{\mathbf{c}_t}^2\mathbf{I})$$

Vaswani *et al.*, **Attention is All You Need**, NIPS'17



Reconstruction Model

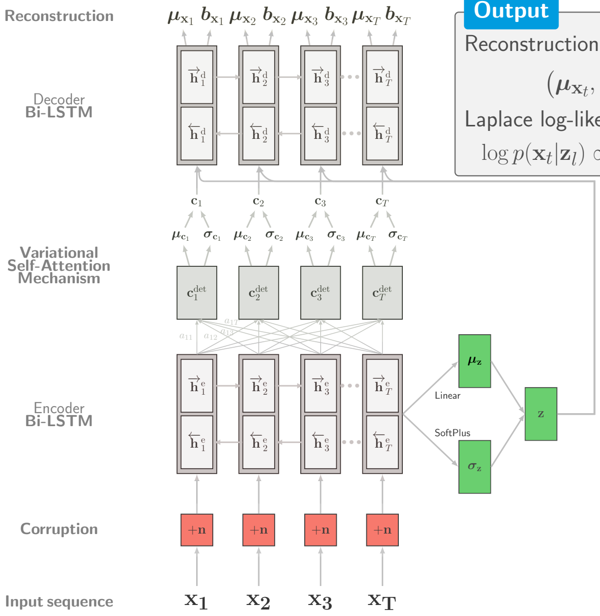


Decoder

Another Bi-LSTM.

$$\text{Decoder} \equiv \text{Bi-LSTM}([\mathbf{z};\mathbf{c}])$$

Reconstruction Model



Output

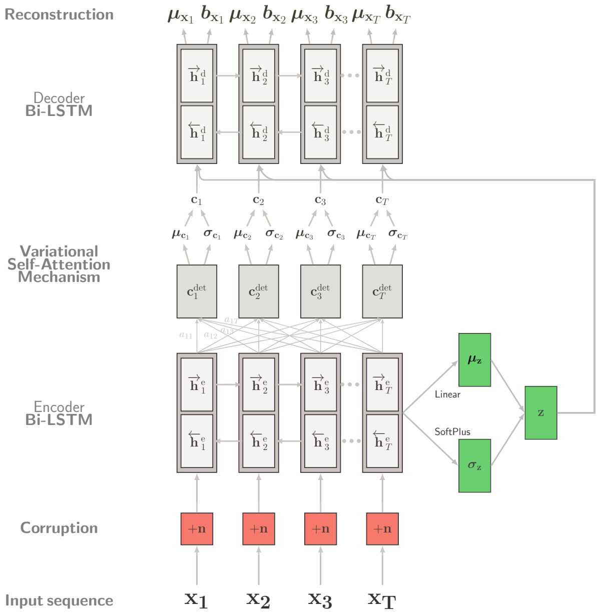
Reconstruction parameters:

$$\left(\boldsymbol{\mu}_{\mathbf{x}_t}, \boldsymbol{b}_{\mathbf{x}_t}\right)_{t=1}^{T}$$

Laplace log-likelihood:

$$\log p(\mathbf{x}_t|\mathbf{z}_l) \propto \|\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x}_t}\|_1$$

Reconstruction Model



Loss Function

Loss Function

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(n)}) = -\mathbb{E}_{\mathbf{z} \sim \tilde{q}_\phi(\mathbf{z}|\mathbf{x}^{(n)}),\ \mathbf{c}_t \sim \tilde{q}^a_\phi(\mathbf{c}_t|\mathbf{x}^{(n)})} \Big[\log p_\theta\big(\mathbf{x}^{(n)}|\mathbf{z}, \mathbf{c}\big) \Big]$$

$$+ \lambda_{\mathrm{KL}} \left[\mathcal{D}_{\mathrm{KL}}\Big(\tilde{q}_\phi(\mathbf{z}|\mathbf{x}^{(n)}) \| p_\theta(\mathbf{z})\Big) + \eta \sum_{t=1}^{T} \mathcal{D}_{\mathrm{KL}}\Big(\tilde{q}^a_\phi(\mathbf{c}_t|\mathbf{x}^{(n)}) \| p_\theta(\mathbf{c}_t)\Big) \right]$$

$\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler Divergence

Loss Function

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(n)}) = \overbrace{-\mathbb{E}_{\mathbf{z}\sim\tilde{q}_\phi(\mathbf{z}|\mathbf{x}^{(n)}),\ \mathbf{c}_t\sim\tilde{q}^a_\phi(\mathbf{c}_t|\mathbf{x}^{(n)})}\Big[\log p_\theta\big(\mathbf{x}^{(n)}|\mathbf{z}, \mathbf{c}\big)\Big]}^{\textbf{Reconstruction Term}}$$

$$+ \lambda_{\mathrm{KL}}\Bigg[\underbrace{\mathcal{D}_{\mathrm{KL}}\Big(\tilde{q}_\phi(\mathbf{z}|\mathbf{x}^{(n)})\|p_\theta(\mathbf{z})\Big)}_{\textbf{Latent Space KL loss}} + \eta\underbrace{\sum_{t=1}^{T}\mathcal{D}_{\mathrm{KL}}\Big(\tilde{q}^a_\phi(\mathbf{c}_t|\mathbf{x}^{(n)})\|p_\theta(\mathbf{c}_t)\Big)}_{\textbf{Attention KL loss}}\Bigg]$$

$\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler Divergence

Training Framework

Optimization & Regularization

- About 270k parameters to optimize
- *AMS-Grad* optimizer[3]
- Xavier weight initialization[4]
- Denoising autoencoding criterion[5]
- Sparse regularization in the encoder Bi-LSTM[6]
- KL cost annealing[7]
- Gradient clipping[8]

Training executed on a single GPU (NVIDIA GTX 1080 TI)

---

[3] Reddi, Kale & Kumar, **On the Convergence of Adam and Beyond**, ICLR'18

[4] Bengio *et al.*, **Understanding the Difficulty of Training Deep Feedforward Neural Networks**, AISTATS'10

[5] Bengio *et al.*, **Denoising Criterion for Variational Auto-Encoding Framework**, AAAI'17

[6] Arpit *et al.*, **Why Regularized Auto-Encoders Learn Sparse Representation?**, ICML'16

[7] Bowman, Vinyals *et al.*, **Generating Sentences from a Continuous Space**, SIGNLL'16

[8] Bengio *et al.*, **On the Difficulty of Training Recurrent Neural Networks**, ICML'13

Proposed Approach

Reconstruction Model

Detection

Proposed Approach

Reconstruction Model

Detection

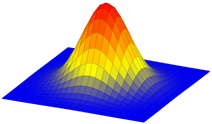
Anomaly Scores

Anomaly Scores

- Use the **reconstruction error** as anomaly score.

$$\textbf{Anomaly Score} = \mathbb{E}_{\mathbf{z}_l \sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\left\| \mathbf{x} - \underbrace{\mathbb{E}[p_\theta(\mathbf{x}|\mathbf{z}_l)]}_{\boldsymbol{\mu}_\mathbf{x}} \right\|_1 \right]$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathrm{Normal}(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\sigma}^2_\mathbf{z}\mathbf{I})$$



$$\mathbf{z}_l \sim q_\phi(\mathbf{z}|\mathbf{x})$$

Anomaly Scores

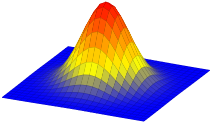
- Use the **reconstruction error** as anomaly score.

$$\textbf{Anomaly Score} = \mathbb{E}_{\mathbf{z}_l \sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\left\| \mathbf{x} - \underbrace{\mathbb{E}[p_\theta(\mathbf{x}|\mathbf{z}_l)]}_{\boldsymbol{\mu}_\mathbf{x}} \right\|_1 \right]$$

- Take the **variability of the reconstructions** into account.

$$\textbf{Anomaly Score} = -\underbrace{\mathbb{E}_{\mathbf{z}_l \sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z}_l) \right]}_{\text{"Reconstruction Probability"}}$$

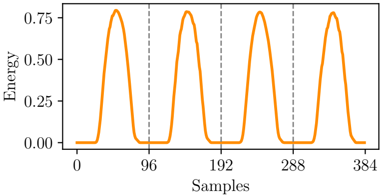
$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathrm{Normal}(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\sigma}_\mathbf{z}^2 \mathbf{I})$$



$$\mathbf{z}_l \sim q_\phi(\mathbf{z}|\mathbf{x})$$

Experiments & Results

Solar PV Generation



(Production in a day without clouds)

- Provided by

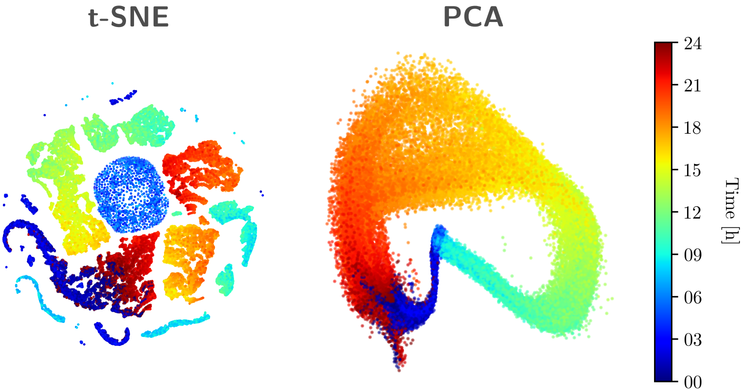
c|side

- Recorded every 15min (96 samples per day)
- Data normalized to the installed capacity
- Daily seasonality
- Unlabelled

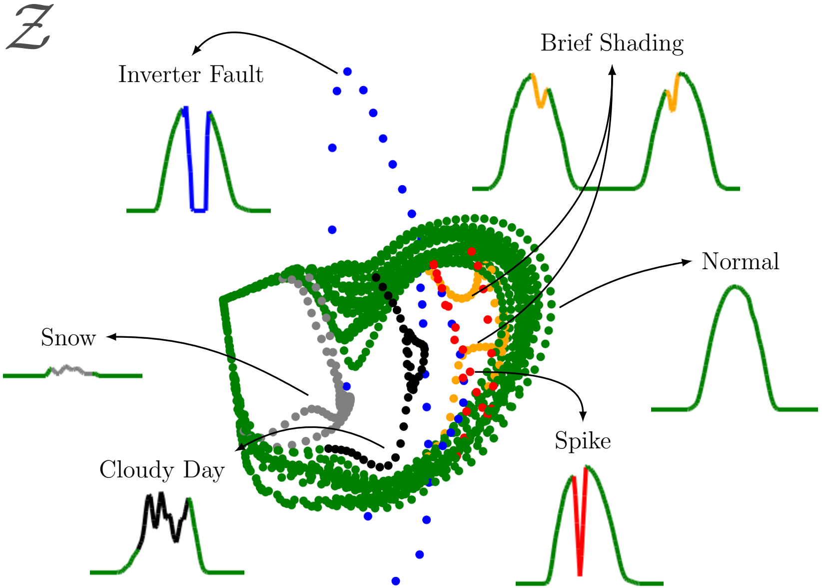
Variational Latent Space

$\mathbf{z}$-space in 2D ($\mathcal{X}_{\text{train}}^{\text{normal}}$)

$T = 32$ (< 96)
$d_{\mathbf{z}} = 3$
online mode

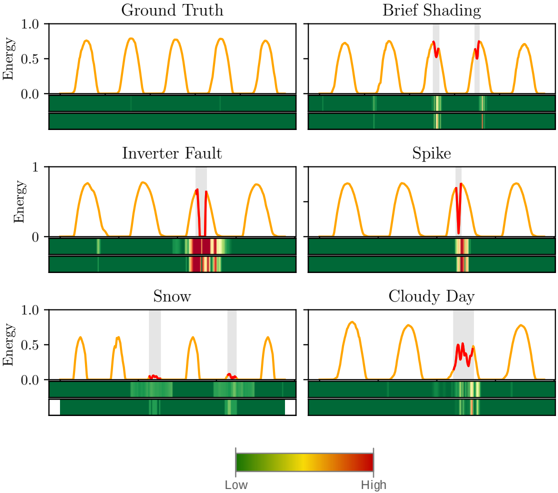


Finding Anomalies in the Latent Space



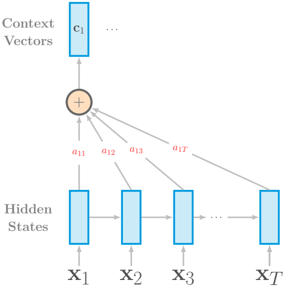
Anomaly Scores

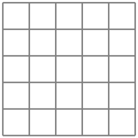
Reconstruction Error (top bar) — **−Reconstruction Probability** (bottom bar)



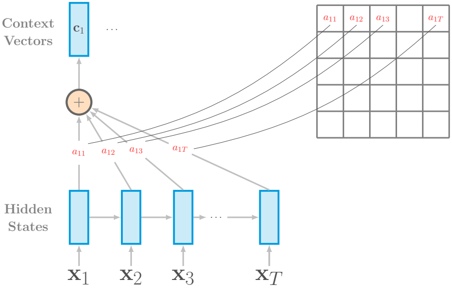
Attention Visualization

Attention Map

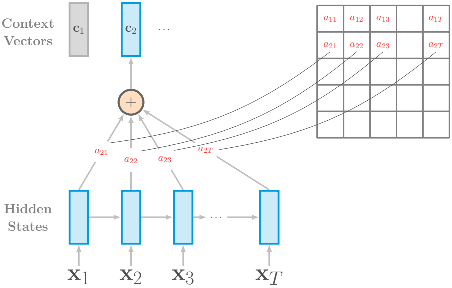




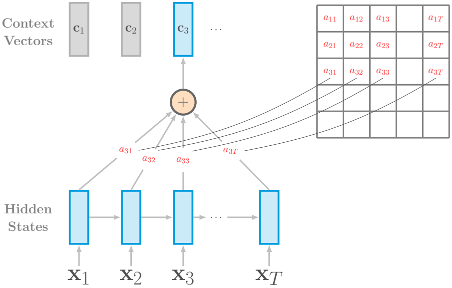
Attention Map



Attention Map

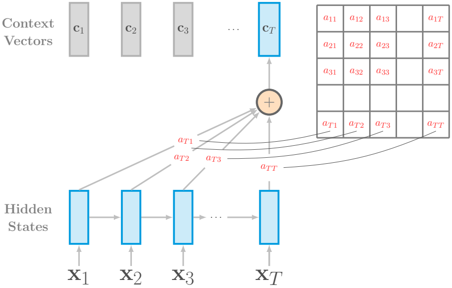


Attention Map



Attention Visualization

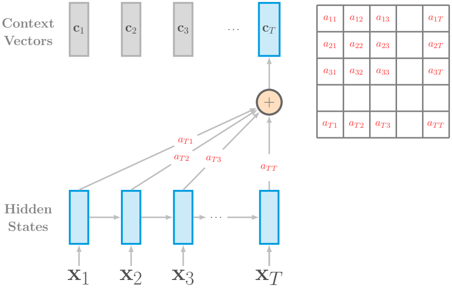
Attention Map



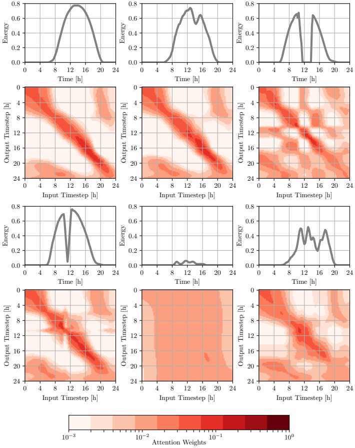
Examples

Attention Visualization

Attention Map



Examples



Conclusions & Future Work

- Effective on detecting anomalies in time series data;
- Unsupervised;
- Suitable for both **univariate and multivariate** data;
- **Efficient**: inference and anomaly scores computation is fast;
- **General**: works with other kinds of sequential data (e.g., **text**, **videos**);
- Exploit the usefulness of the **attention maps** for detection;
- Make it robust to changes of the normal pattern over time.

Acknowledgements

FCT
20 anos
Fundação para a Ciência e a Tecnologia

Institute for Systems and Robotics
Lisboa

SIPG
SIGNAL AND IMAGE PROCESSING GROUP

TÉCNICO LISBOA

Thank you for your attention!

mail@joao-pereira.pt
www.joao-pereira.pt