Representation Learning for Sequence Data with Deep Autoencoding Predictive Components

JUNWEN BAI, WEIRAN WANG, YINGBO ZHOU, CAIMING XIONG

ICLR 2021

Motivation

- Self-supervised learning is trending in sequence representation learning field

- Keys
  - the structure imposed on the latent sequential space
  - information retained in the representations

- Our goals
  - learn latent sequential representations that exhibit a simple structure
  - And retain useful information from inputs

DAPC

- We propose deep autoencoding predictive components (DAPC) learning

- Major take-aways
  - DAPC models mutual information between past and future (predictive information or PI)
  - PI estimation is exact under Gaussian assumption (simple structure)
  - Learning of DAPC is regularized by masked reconstruction (input information)

Method

- Given a time series data $X = \{x_1, x_2, \ldots\}$ where $x_t \in R^n$ and its corresponding latent sequence $Z = \{z_1, z_2, \ldots\}$ where $z_t \in R^d$
- $Z_{past} = (z_{-T+1}, \ldots, z_0)$ and $Z_{future} = (z_1, \ldots, z_T)$

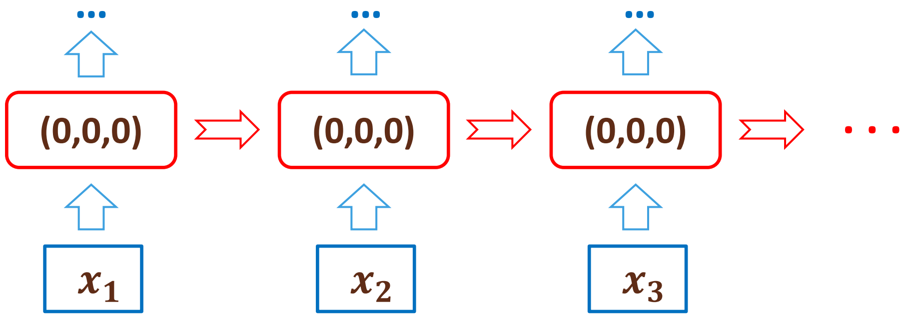
- Predictive information (PI) is the MI between Z_{past} and Z_{future}
  - $MI(Z_{past}, Z_{future}) = H(Z_{past}) + H(Z_{future}) - H(Z_{past}, Z_{future})$

- $MI(Z_{past}, Z_{future}) = \ln|\Sigma_T(Z)| - \frac{1}{2}\ln|\Sigma_{2T}(Z)|$
  - Gaussian assumption
  - $\Sigma_T(Z)$ is the covariance matrix of the length-T Gaussian distribution

Method

- Encode x_i to z_i through neural networks (RNN, Transformer, …)

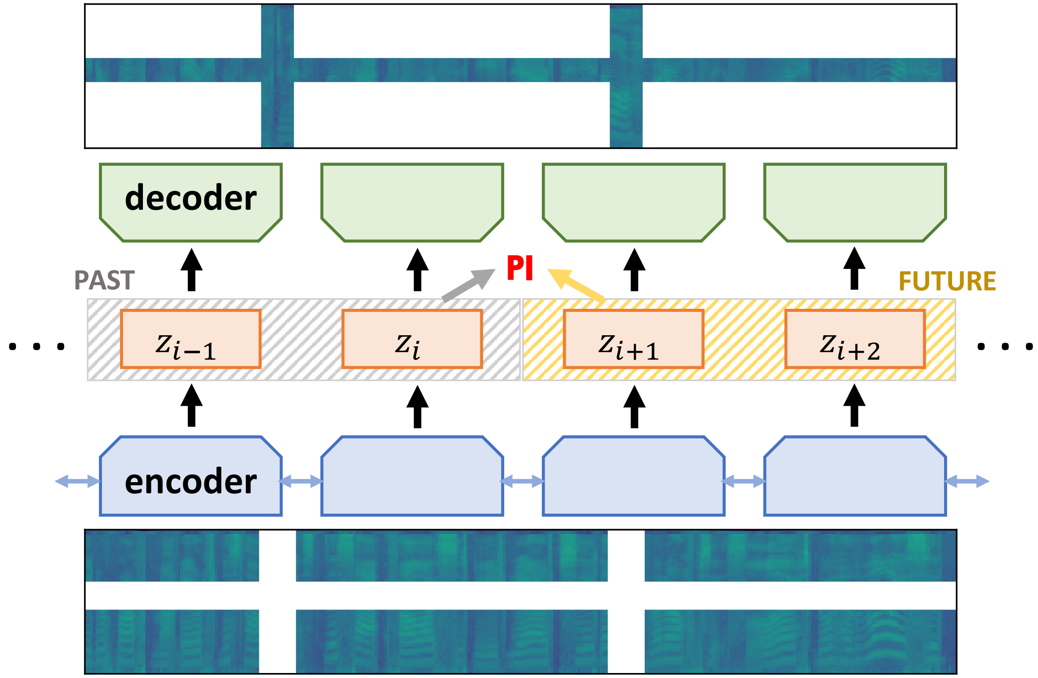
- If we only use PI as objective, powerful neural networks might learn trivial latent codes (see below) since this would give very high PI



- Our solution: regularize the learning with masked reconstruction

Method

- Masked Reconstruction
  - encoder: $e(\cdot)$
  - decoder: $g(\cdot)$
  - mask inputs
  - reconstruct the masked portion
- $R = \left\| (1 - M) \odot \left(X - g\left(e(X \odot M)\right)\right) \right\|_2^2$

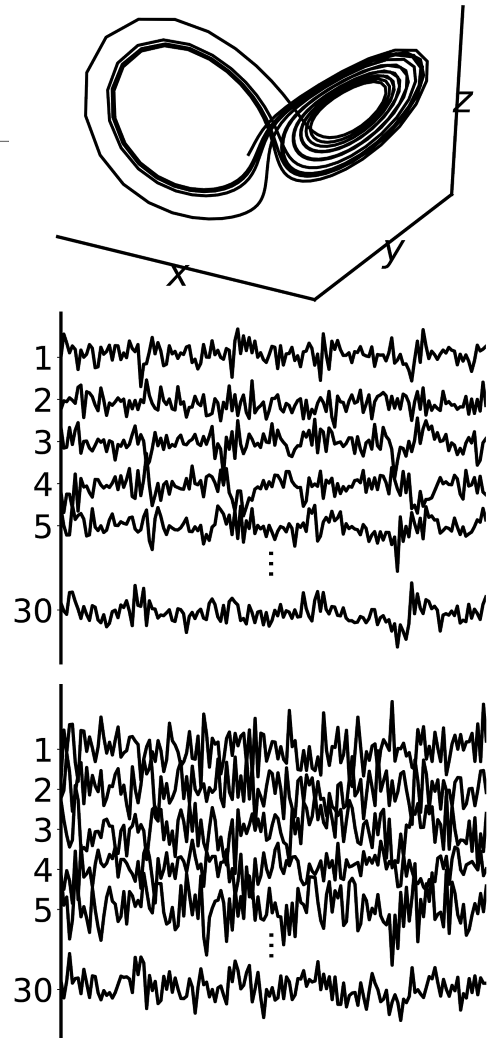


Lorenz Attractor

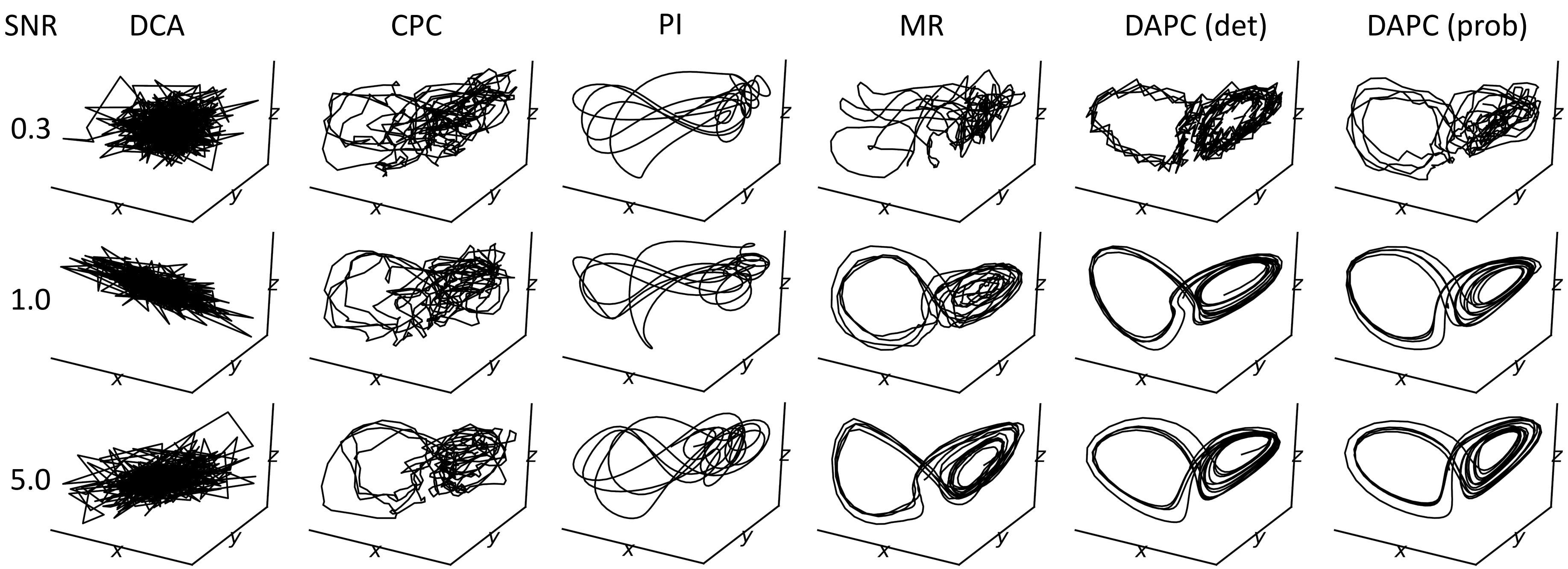
- Lorenz Attractor is 3-d time series data

- We non-linearly lift the 3-d data to 30-d data with a random NN of 2 hidden layers and further add white noise

- Task: recover the 3-d ground-truth Lorenz Attractor from the 30-d data without any supervision



Lorenz Attractor



- Other domains with the same setup: weather, biology, etc.

Automatic Speech Recognition

- To show DAPC is scalable, we apply it on 2 speech corpus, WSJ and LibriSpeech
- Competitive results compared with other SOTA representation learning methods

Methods	*dev93*	*eval92*
Finetune on 15 hours		
w.o. pretrain	12.91	8.98
DAPC	12.31	7.74
DAPC + multi-scale PI	12.15	
DAPC + shifted recon	11.93	
DAPC + both	**11.57**	**7.34**

Methods	WER (%)
wav2vec [8]	6.92
discrete BERT+vq-wav2vec [38]	4.5
wav2vec 2.0 [12]	**2.3**
DeCoAR [43]	6.10
TERA-large [44]	5.80
w.o. pretrain	5.11
MR	5.02
DAPC	4.86
DAPC+multi-scale PI+shifted recon	**4.70**

Q & A

Thanks for listening!