Continuity-Preserving Convolutional Autoencoders for Learning Continuous Latent Dynamical Models from Images

Aiqing Zhu, Yuting Pan, Qianxiao Li *

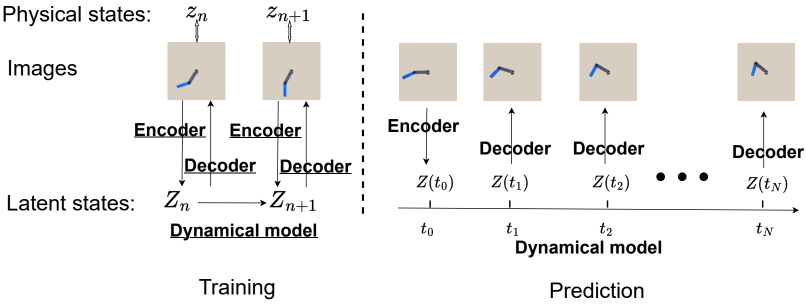
Department of Mathematics
National University of Singapore

Main contributation

- We propose a mathematical formulation for learning continuous dynamics from image data to describe the continuity of latent states.
- We demonstrate that the latent states will evolve continuously with the underlying dynamics if the filters are Lipschitz continuous.
- We introduce a regularizer to promote the continuity of filters and, consequently, preserve the continuity of the latent states.
- We perform several experiments across various scenarios to verify the effectiveness of the proposed method.

Learning latent dynamical models from image

$$\dot{z} = f(z), \quad z \in \mathcal{Z} \subset \mathbb{R}^D,$$

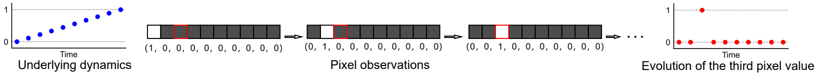


Our goal:

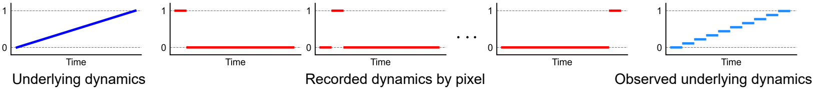
- **learn an encoder that extract latent states consistent with the assumed latent dynamical system;**
- discover a dynamical model that accurately captures the underlying latent dynamics;
- identify a decoder capable of reconstructing the pixel observations.

Discrete nature of pixel coordinates

Pixels do not align with the continuous evolution of underlying dynamics

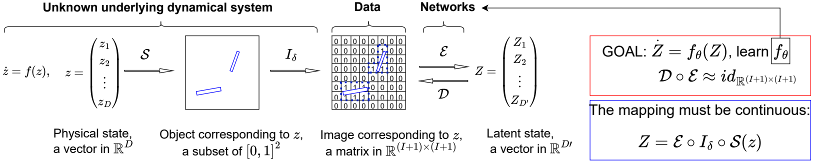


A single pixel white square, initially located at the leftmost position, moves uniformly to the right against a black background (plotted in gray for clarity).



We further assume that the object occupies a very small volume and its motion is recorded in continuous time periods. Then variations smaller than δ can not be captured in the image.

Mathematical formulation


Unknown underlying dynamical system
Data
Networks
$\dot{z} = f(z)$, $z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_D \end{pmatrix}$
$\mathcal{S}$
I_δ
$\mathcal{E}$
$\mathcal{D}$
$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{D'} \end{pmatrix}$
Physical state, a vector in $\mathbb{R}^D$
Object corresponding to z, a subset of $[0,1]^2$
Image corresponding to z, a matrix in $\mathbb{R}^{(I+1)\times(I+1)}$
Latent state, a vector in $\mathbb{R}^{D'}$

GOAL: $\dot{Z} = f_\theta(Z)$, learn $\boxed{f_\theta}$

$\mathcal{D} \circ \mathcal{E} \approx id_{\mathbb{R}^{(I+1)\times(I+1)}}$

The mapping must be continuous:

$$Z = \mathcal{E} \circ I_\delta \circ \mathcal{S}(z)$$

Definition. A sequence of functions $\{g_\delta(z) : \mathcal{Z} \to \mathbb{R}^d | \delta \in \{1/(I+1)\}_{I=1}^\infty\}$ is called δ-continuous if there exists a constant c_g such that for all $z_1, z_2 \in \mathcal{Z}$, there exists a δ^* such that if $\delta \leq \delta^*$, then $\|g_\delta(z_1) - g_\delta(z_2)\| \leq c_g \|z_1 - z_2\|$.

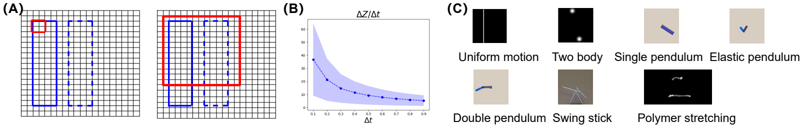
Therefore, our objectives for the autoencoder are as follows:

- **Find the encoders $\{\mathcal{E}_\delta\}$ such that $\mathcal{E}_\delta \circ \boldsymbol{I}_\delta \circ \mathcal{S}(z)$ is δ-continuous.**
- Find the decoders $\{\mathcal{D}_\delta\}$ such that $\mathcal{D}_\delta \circ \mathcal{E}_\delta$ is approximately the identity mapping.

If it can be achieved, with an additional assumption that the pixel size δ^* of the image is sufficiently small, then there exists a constant $c_\mathcal{E}$ such that

$$\|Z_n - Z_{n+1}\| = \|\mathcal{E}_{\delta^*} \circ \boldsymbol{I}_{\delta^*} \circ \mathcal{S}(z_n) - \mathcal{E}_{\delta^*} \circ \boldsymbol{I}_{\delta^*} \circ \mathcal{S}(z_{n+1})\| \leq c_\mathcal{E} M_f \Delta t.$$

Why standard CNN autoencoders fail?



(A) Illustration of convolution operation. The red boxes represent the filter of size $\mathcal{O}(1)$ or $\mathcal{O}(1/\delta)$. The blue box represents the object. The solid line indicates its initial position, while the dashed line represents its position after motion. **(B)** The variation of latent states divided by Δt for the two-body system, where the encoder is a one-layer CNN with parameters uniformly sampled from $[-1, 1]$. The shaded region represents one standard deviation. **(C)** Examples of motion where the positions of the objects after variation only partially overlap with their positions before variation.

Quantifying continuity of CNN autoencoders

Theorem. Assume that the underlying dynamical system is a rigid body motion on a two-dimensional plane. Let $c_{\mathcal{W}}$ be constants satisfying

$$\max_{l=1,\cdots,L^*} |\mathcal{W}_l(y_1) - \mathcal{W}_l(y_2)| \leq c_{\mathcal{W}} \|y_1 - y_2\|,$$

and if $s_l = 2$ for $l = 1, \cdots, L^* - 1$, then for any $z_1 = (z_1^t, z_1^r), z_2 = (z_2^t, z_2^r) \in \mathcal{Z}$,

$$\begin{aligned} &\|\mathcal{E}_\delta \circ \boldsymbol{I}_\delta \circ \mathcal{S}(z_1) - \mathcal{E}_\delta \circ \boldsymbol{I}_\delta \circ \mathcal{S}(z_2)\| \\ &\leq C c_{\mathcal{W}} \|z_1^r - z_2^r\| + \frac{C c_{\mathcal{W}}}{2^{L^*-1}} \left\|z_1^t - z_2^t\right\|, \quad \text{as} \quad \delta \to 0. \end{aligned}$$

Here C is a constant independent of δ and z.

Function $\mathcal{W}_l : \mathbb{R}^2 \to [-1, 1]$ is independent of δ and represent standard CNN filter $\boldsymbol{W}_l^\delta$:

$$[\boldsymbol{W}_l^\delta]_{j_1,j_2} = \mathcal{W}_l(j_1\delta, j_2\delta)\varepsilon_l, \quad \text{where } \mathcal{W}_l(x) = 0 \text{ if } x \notin [0, J_l\delta]^2.$$

Method to preserve continuity of CNN autoencoders

Note the fact that $\max_l \max_{j_1,j_2} |\mathcal{W}_l(j_1\delta, j_2\delta)|/(\lceil J_l/2\rceil\delta) \leq c_{\mathcal{W}}$, larger filters are necessary to ensure continuity.

We recommend using the nonlocal operators method, an image processing technique that promotes image continuity. This approach requires only the following regularizer for the filters:

$$\mathcal{J} = \lambda_J \sum_{l=1}^{L^*} \sum_{i_1,i_2,j_1,j_2=-\hat{J}}^{J^l+\hat{J}} (W^l_{i_1,i_2} - W^l_{j_1,j_2})^2 k\left((i_1\delta, i_2\delta), (j_1\delta, j_2\delta)\right).$$

Experiments

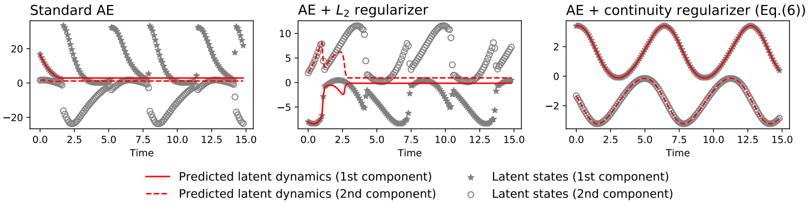
Quantitative results

Method / Dataset	CpAE		Hybrid scheme		AE +Neural ODE		AE+HNN		AE+SympNet	
	VPT	VPF	VPT	VPF	VPT	VPF	VPT	VPF	VPT	VPF
Damped pendulum	99.2±8.5	99.2	95.4±15.0	88.3	50.7±31.2	23.3	—	—	—	—
Elastic pendulum	72.1±27.2	36.7	49.5±24.2	10.0	30.6±18.5	1.7	—	—	—	—
Double pendulum	69.1±31.5	40.0	46.8±21.4	4.6	24.3±13.8	0.0	11.0±4.2	0.0	15.1±12.8	0.0
Swing stick	57.4±20.4	11.1	13.7±5.1	0.0	14.4±7.5	0.0	24.1±14.5	0.0	14.8±12.2	0.0

$$\mathrm{VPT} = \arg\max_t \{t \le T \mid \mathrm{PMSE}(X_\tau, \bar{X}_\tau) \le \varepsilon, \forall \tau \le t\},$$

VPF represents the frequency of test trajectories for which VPT $= 1$.

Continuity of latent states



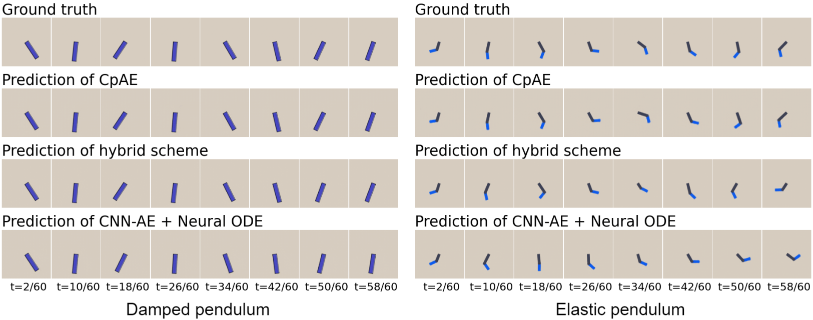
Standard AE

AE + L_2 regularizer

AE + continuity regularizer (Eq.(6))

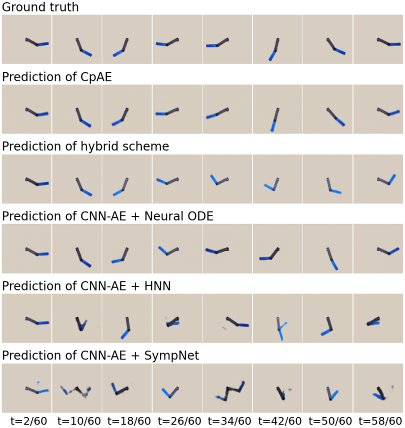
Experiments

Predictions for simulation data

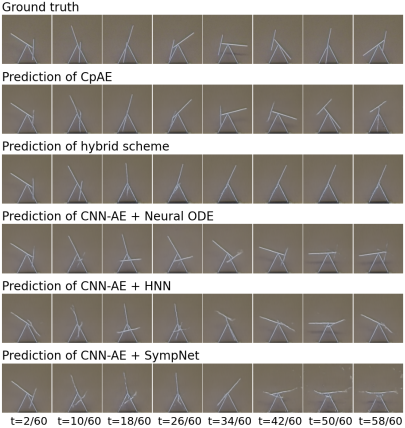


Experiments

Predictions for real-world data


Ground truth
Prediction of CpAE
Prediction of hybrid scheme
Prediction of CNN-AE + Neural ODE
Prediction of CNN-AE + HNN
Prediction of CNN-AE + SympNet
t=2/60
t=10/60
t=18/60
t=26/60
t=34/60
t=42/60
t=50/60
t=58/60

Double pendulum


Ground truth
Prediction of CpAE
Prediction of hybrid scheme
Prediction of CNN-AE + Neural ODE
Prediction of CNN-AE + HNN
Prediction of CNN-AE + SympNet
t=2/60
t=10/60
t=18/60
t=26/60
t=34/60
t=42/60
t=50/60
t=58/60

Swing stick

Thanks for your attention!