Biologically-Inspired Continual Learning of Human Motion Sequences

Joachim Ott, Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

Problem Statement

Common machine learning models are not able to retain or transfer knowledge as biological brains can, since they suffer from catastrophic forgetting. Continual learning explores solutions to overcome these limitations, for example with generative replay.

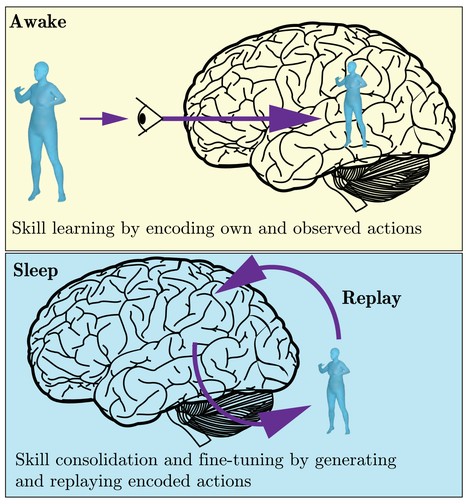
In most studies that use a generative replay approach, the generated samples are just a tool to maintain classification accuracy. In contrast, good generative performance, e.g., in imitating or reproducing motions, is very important in many real life situations. This biology rooted feature of continual learning to generate is still under-explored in machine learning.

Goal

We research how a Biologically-Inspired Conditional Temporal Variational Autoencoder (BI-CTVAE) performs in a novel continual-learning-to-generate (CL2Gen) setting.

Background

Humans show an amazing ability to learn continuously without forgetting; old knowledge is retained and can be used in new situations. In training for a sport like tennis, humans learn and refine their strokes not only through practice, but also through mental rehearsal and observation.
To consolidate the acquired procedural knowledge, motion representations are regenerated or replayed during sleep. This knowledge can then be consolidated into long-term memory in the hippocampus.



Proposed BI-CTVAE Model

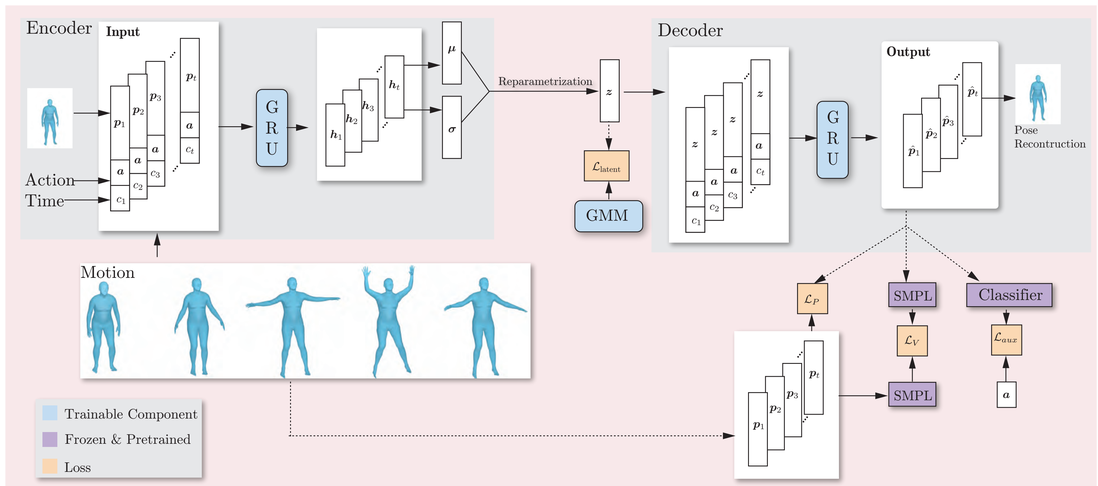


Fig. 1. Overview of BI-CTVAE components.
Encoder: Input is a motion sequence of poses p1...pT , each concatenated with the action label a and a time index c1...cT . The input is first processed by a multi-layered GRU network. The hidden state of the last layer is used to calculate the latent vectors, μ and σ. These are used to sample a motion latent representation z. **Decoder**: Input is a sequence of repeated z, each concatenated with action label a and time index c1...cT . Decoder output is the reconstructed motion. GMM: The per-class GMM components allow sampling of classes learned in previous tasks.

Motion representation: The full pose frame p consists of the body pose r with 23 joints + global rotation and the displacement of the root joint. The 23 joint rotations and one global rotation are transformed into a 6D matrix representation.

Loss:

$\mathcal{L}_V$ Vertex loss uses a pretrained and frozen SMPL model [1] and calculates an L2 loss for every pose vertex in the sequence.

$\mathcal{L}_P$ Reconstruction loss uses the L2 loss on the reconstructed pose.

$\mathcal{L}_{aux}$ Auxiliary loss is the cross entropy loss calculated using the predicted label obtained from a pretrained and frozen classifier.

The standard regularization term for training VAEs is the Kullback Leibler Divergence (KL) loss, since we use a separate mode for every class, we modify the regularization term as follows:

$$\mathcal{L}_{\text{latent}}(\boldsymbol{M}, a, \boldsymbol{\phi}, \mathcal{X}) = \frac{1}{2}\sum_{j=1}^{256}\left(1+\log({\sigma_j^{(M)}}^2)-\log({\sigma_j^{a}}^2)-\frac{\left(\mu_j^{(M)}-\mu_j^a\right)^2+{\sigma_j^{(M)}}^2}{{\sigma_j^{a}}^2}\right)$$

where $\boldsymbol{M}$ is the input motion sequence, a the class label, $\mathcal{X}$ is the set of trainable means and standard deviations, $\boldsymbol{\mu}^{(M)}$ and $\boldsymbol{\sigma}^{(M)}$ are calculated from $\boldsymbol{M}$ by the model with parameters $\boldsymbol{\phi}$, $\boldsymbol{\mu}^a$ and $\boldsymbol{\sigma}^a$ are the trainable mean and standard deviation of the mode corresponding to class a, μ_j^a and σ_j^a are jth elements respectively of $\boldsymbol{\mu}^a$ and $\boldsymbol{\sigma}^a$.

Final loss consists of the loss components above and scaling factors:

$$\mathcal{L} = \mathcal{L}_V + \mathcal{L}_P + \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} + \lambda_{aux}\mathcal{L}_{aux}$$

CL2Gen: The model is trained incrementally on tasks where each task is represented by a set of action classes. The goal is for the model to maintain the ability to generate representative samples of classes of previous tasks even after training on the new classes within each new task.

Experiments and Results

The HumanAct12 human motion dataset is split into 6 tasks with 2 classes each. We then train for CL2Gen and evaluate performance with common metrics for generative motion models:

Model	Replay	Samples*	Accuracy↑	FID↓	Diversity→	Multimodality→
Ground Truth [23]	No	-	$99.7^{\pm 0.01}$	$0.092^{\pm 0.007}$	$6.85^{\pm 0.05}$	$2.45^{\pm 0.04}$
GRU (offline)	No	-	$74.2^{\pm 1.04}$	$0.71^{\pm 0.00}$	$6.63^{\pm 0.02}$	$4.31^{\pm 0.04}$
Ours (offline)	No	-	$84.0^{\pm 1.37}$	$0.89^{\pm 0.01}$	$6.69^{\pm 0.02}$	$3.48^{\pm 0.02}$
GRU	No	-	$13.1^{\pm 1.38}$	$11.18^{\pm 0.24}$	$5.91^{\pm 0.05}$	$5.47^{\pm 0.05}$
BI-CTVAE	No	-	$14.8^{\pm 1.00}$	$13.18^{\pm 0.24}$	$5.22^{\pm 0.02}$	$4.50^{\pm 0.04}$
	1/16	Real	$42.2^{\pm 2.43}$	$\mathbf{1.73^{\pm 0.03}}$	$\mathbf{6.60^{\pm 0.03}}$	$4.72^{\pm 0.03}$
	1/16	Gen	$43.1^{\pm 4.62}$	$2.31^{\pm 0.02}$	$6.55^{\pm 0.03}$	$4.48^{\pm 0.02}$
	1/5	Gen	$48.9^{\pm 2.49}$	$2.21^{\pm 0.02}$	$6.54^{\pm 0.02}$	$4.27^{\pm 0.02}$
GRU aux	1/16	Gen	$21.9^{\pm 1.58}$	$7.25^{\pm 0.16}$	$6.06^{\pm 0.02}$	$5.34^{\pm 0.03}$
	1/5	Gen	$42.5^{\pm 2.83}$	$5.52^{\pm 0.13}$	$6.06^{\pm 0.03}$	$5.12^{\pm 0.03}$
BI-CTVAE aux	1/16	Real	$37.5^{\pm 2.74}$	$3.33^{\pm 0.06}$	$6.75^{\pm 0.02}$	$5.29^{\pm 0.07}$
	1/16	Gen	$61.6^{\pm 2.37}$	$3.67^{\pm 0.04}$	$6.56^{\pm 0.02}$	$4.15^{\pm 0.01}$
	1/5	Gen	$\mathbf{78.6^{\pm 1.40}}$	$2.70^{\pm 0.01}$	$6.57^{\pm 0.03}$	$\mathbf{3.07^{\pm 0.03}}$

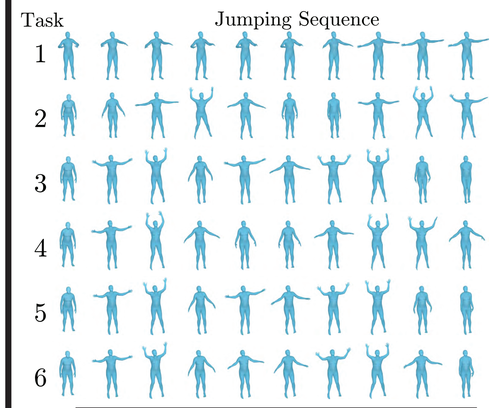


Fig. 2. Generated 'jump' action samples of the model trained with replay 1/5 and auxiliary loss. After training on 'jump' in task 2, the model can generate this action.
Even after training on the next tasks 3 to 6 with new action classes, the model retains its ability to generate realistic sequences for 'jump'.

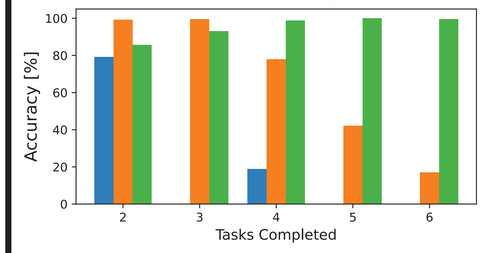


Fig. 3. Classification accuracy of generated 'jump' samples after each task.
Only the 1/5 replay model maintains performance after training on subsequent tasks.

Conclusion

The final classification accuracy of BI-CTVAE on the HumanAct12 dataset after sequentially learning all action classes is 78%, which is 63% higher than using no-replay, and only 5.4% lower than a state-of-the-art offline trained GRU model.

References and Sources
Bain Shape: Warren H Lewis and Henry Gray. Anatomy of the human body. Lea & Febiger, Philadelphia doi, 10, 1918.
[1] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J., "Smpl: A skinned multi-person linear model," ACM transactions on graphics (TOG), vol. 34, no. 6, pp. 1–16, 2015.

Part of the work was done at Starmind AG by the first author. The subsequent work at University of Zurich was partially supported by the Swiss National Science Foundation CA-DNNEdge (208227).