Multi-Modal Representation Learning for Mapping Between Body Motion and Visual Imagery

Tobias Jülg[1,2], Florian Walter[2], Dongmin Kim[1], Hoshinori Kanazawa[1], Alois Knoll[2], Yasuo Kuniyoshi[1]
[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]School of Computation, Information and Technology, Technical University of Munich, Germany
{*juelg, walter, knoll*}*@cit.tum.de*, {*d-kim, kanazawa, kuniyosh*}*@isi.imi.i.u-tokyo.ac.jp*

1 Introduction

Human infants need to learn autonomously to control their body in order to achieve higher level tasks. For this, it is crucial to have an internal representation of the body and environment states. An important requirement is that it should be multi-modal, to learn a bidirectional mapping between body states, represented as mental images, and related joint angles.

In this work, we model this mapping in a novel way by combining two basic tools from representation learning: We use a multi-modal self-organizing map (SOM) [1] to learn an internal representation for vision and proprioception of a simulated infant. As we will show, adding an autoencoder (AE) represents the visual stimuli in a low-dimensional latent space, reduces the computational complexity and increases the quality of the learned representation.

2 Method

2.1 Dataset

We have created a dataset with $N = 20000$ samples generated with the infant simulation from [2]. It contains data from two modalities for training the AE and the multi-modal SOM. Each sample in the dataset corresponds to an arm pose of the simulated infant and contains data from vision (modality m_1) and proprioception (modality m_2):

$$\mathscr{D} = \{(d_i^{(m_1)}, d_i^{(m_2)})\}_{i=1}^{N} \tag{1}$$

The experiment setup and a sample can be seen in Figure 1. For the vision modality, we use a camera which is attached to the position of the infant's right eye. It can take wide angle RGB images of 300 by 500 pixels. For the proprioception modality the five joint angles of the right arm are obtained.

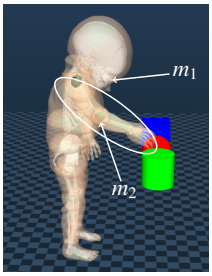
The poses used in the dataset were sampled randomly in joint space. Samples where the arm was not visible were discarded. For training we used a 60/20/20 split.

2.2 Multi-Modal SOM

The multi-modal SOM for learning a representation of the infants's arm pose builds upon the Multi-Modal Convergence Maps framework from [1].

We use x and y to refer to the map dimensions and m to refer to the modality dimension. We further define the weights of the map $w_{x,y}^{m}$ for the neuron $x, y \in \mathbb{N}$ and modality $m \in \mathscr{M}$.

In order to train the map for a given training sample $d_i \in \mathscr{D}$,



(a) Experiment setup

$$m_1 = \quad \in \mathbb{R}^{300\times500\times3}$$

$$m_2 = \begin{pmatrix} 1.22 \\ 0.60 \\ -0.12 \\ 0.61 \\ 0.91 \end{pmatrix} \in \mathbb{R}^5$$

(b) Sample

Figure 1: (a) shows the experiment setup in the MuJoCo simulation with the infant performing a valid pose and (b) shows the corresponding dataset sample.

one first needs to calculate the neuron activity per modality as the weighted MAE between the current learned neuron representation and the sample. This is done for all modalities separately:

$$a_{x,y}^{(m)} = -\frac{c^{(m)}}{n^{(m)}} \cdot \sum_{j=1}^{n^{(m)}} |w_{x,y}^{(m)} - d_i^{(m)}|_j \quad \forall x, y \; \forall m \in \mathscr{M} \tag{2}$$

$c^{(m)}$ is a factor which attributes each modality a certain influence during training and $n^{(m)}$ is the flattened dimensionality of modality m. Afterwards, one calculates the combined neuron activity over all modalities:

$$A_{x,y} = \frac{\sum_{m\in\mathscr{M}} a_{x,y}^{(m)}}{\sum_{m\in\mathscr{M}} c^{(m)}} \tag{3}$$

The neuron with the highest activity, i.e. the one that most closely represents the training sample, is picked as the winning neuron: $x_{\text{win}}, y_{\text{win}}$. The weight update is then calculated by the error, a learning rate λ and an exponentially decaying neighborhood function $\mu : \mathbb{N}^4 \to \mathbb{R}_{\geq 0}$ which has the highest value at the winning neuron:

$$dw_{x,y}^{(m)} = (w_{x,y}^{(m)} - d_i^{(m)}) \cdot \lambda \cdot \mu(x, y, x_{\text{win}}, y_{\text{win}}) \tag{4}$$

We extended the weight update proposed in [1] to support mini batches by averaging over individual updates:

$$dw_{x,y}^{(m)} = \frac{1}{|B_i|} \sum_{s\in B_i} dw_{x,y,s}^{(m)} \tag{5}$$

where B_i is the ith batch and $dw_{x,y,s}^{(m)}$ is the update calculated by the sth item of this batch for modality m.

After training a multi-modal SOM, it is possible to map between its modalities. Given, for example, a sample $s^{(m_2)}$ where only m_2 is known, then $\hat{s}^{(m_1)}$ can be predicted by simply calculating the activity for $s^{(m_2)}$ and returning the m_1 component of the winning neuron: $w^{(m_1)}_{x_{win},y_{win}}$.

2.3 Architecture

Our SOM with pre-trained AE (SOM-AE) architecture is summarized in Figure 2. Visual input is represented in a latent space of dimension 8 by an AE that was trained in advance. It consists out of an encoder $e : \mathbb{R}^{64\times64} \to \mathbb{R}^8$ (dimensions result from pre-processing) and a decoder $d : \mathbb{R}^8 \to \mathbb{R}^{64\times64}$ which are trained on the MSE reconstruction loss. Thus, the training input of the SOM for a sample $d_i^{(m_1)}$ becomes $l_i^{(m_1)} = e(d_i^{(m_1)})$. This reduces the map size from 10.5GB to only 33.3MB. Note, that it is possible to cache the encodings offline, thus avoiding a new bottleneck.

We argue that this combination of SOMs and AEs is biologically plausible, as it can be compared to the hierarchical structure of visual processing in the human brain. The AE uses a convolutional neural network architecture that has been argued to share similarities with the visual cortex [3]. Thus, the latent space can be compared to the highest level of abstraction which is fed to multisensory areas which are modeled by the multi-modal SOM.

Combining SOMs and AEs has been studied in a few ways before. The approach which might come closest to ours is Deep Neural Maps [4] which learn both AE and SOM simultaneously, resulting in improved embeddings on unimodal image focused benchmark datasets. In contrast, we first train the AE separately and use the learned latent space to train a biologically plausible multi-modal SOM [1]. We also use a complex dataset composed of different modalities of arm postures from a simulated embodied infant.

3 Results and Conclusion

The performance comparison between the plain SOM and SOM-AE can be seen in Table 1. MAE losses and standard deviations for four mapping scenarios over the test set are provided in the lower part. In the case of SOM-AE, for the vision modality, the decoded output $d(w^{(m_1)}_{x_{win},y_{win}})$ is used for the error calculation with the sample's image $d_i^{(m_1)}$. SOM-AE uses much less computational resources for training, even when the training time of the AE is also included. The performance does not significantly reduce and even increases in some cases. SOM-AE also performs better in terms of learned map quality, as the randomly chosen validation samples in Table 2 suggest.

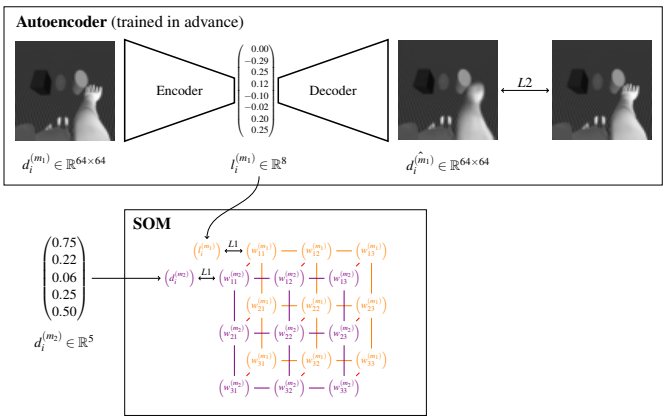


Figure 2: SOM-AE architecture, the visual modality m_1 is first used to train an AE with a latent space of size 8. Afterwards the latent space is used for the multi-modal SOM training.

Given the trained SOM-AE, it is possible to let an agent, which embodies the infant, "imagine" a situation and let it "remember" how to get to it by mapping between modalities, as described earlier.

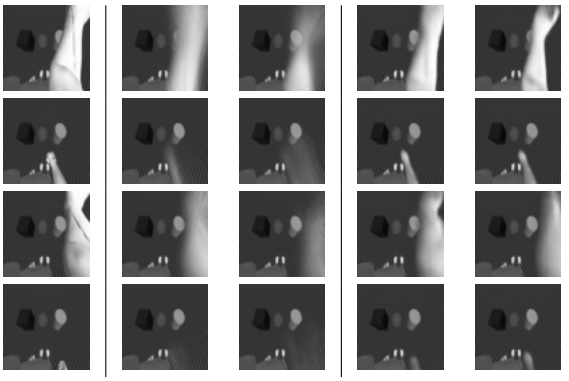
In conclusion, SOM-AE not only significantly reduces the computational requirements to train a multi-modal SOM model, but also improves the qualitative mapping results while keeping the quantitative loss at the same level. Therefore, this framework is a step towards better multi-modal representations and bidirectional modality mapping. As argued before, it can partly be mapped to cognitive perception processes in the human brain and it, thus, models the emergence of interconnection between body motion and visual imagery in an infant's developmental process.

Table 1: Quantitative performance comparison

	SOM	SOM-AE
Batch size	1	**10**
Epochs	200	200
SOM training time	61.5 h	**15.5 min**
AE training time	**0**	9.5 h
GPU	A100 SXM4	**RTX 2080**
VRAM used	42GB	**2.6GB**
Overall test loss	0.068±0.077	**0.058**±0.065
Test loss: $m_1 \to m_1$	0.025±0.020	**0.021**±0.020
Test loss: $m_1 \to m_2$	0.163±0.096	**0.127**±0.087
Test loss: $m_2 \to m_1$	0.046±0.038	**0.041**±0.041
Test loss: $m_2 \to m_2$	**0.039**±0.021	0.042±0.023

Table 2: Qualitative inference comparison

$d_i^{m_1}$	SOM		SOM-AE	
	$m_1 \to m_1$	$m_2 \to m_1$	$m_1 \to m_1$	$m_2 \to m_1$



References

[1] S. Lallee and P. F. Dominey, "Multi-modal convergence maps: from body schema and self-representation to mental imagery," *Adaptive Behavior*, vol. 21, no. 4, pp. 274–285, Aug. 2013.

[2] D. Kim, H. Kanazawa, and Y. Kuniyoshi, "Simulating a human fetus in soft uterus," in *2022 IEEE International Conference on Development and Learning (ICDL)*, 2022, pp. 135–141.

[3] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[4] M. Pesteie, P. Abolmaesumi, and R. Rohling, "Deep neural maps," 2018. [Online]. Available: https://openreview.net/forum?id=HyG76D1wf