Learning Cognitive State Representations from Neuronal and Behavioural Data

Akshey Kumar (akshey.kumar@univie.ac.at)
Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna, Austria
UniVie Doctoral School Computer Science DoCS

Aditya Gilra (aditya.gilra@cwi.nl)
Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Moritz Grosse-Wentrup (moritz.grosse-wentrup@univie.ac.at)
Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna, Austria
Vienna Cognitive Science Hub, University of Vienna, Austria
Research Platform Data Science @ Uni Vienna

Abstract

Explaining how neuronal activity gives rise to behaviour and cognition is a central goal of cognitive neuroscience. With the proliferation of larger neuronal datasets, there have been various attempts to abstract representations of the neuronal data. Some methods consider behavioural decoding to be important while other unsupervised methods like PCA and autoencoder disregard behaviour altogether. Here, we propose an architecture to learn cognitive state representations which preserve information of both the dynamics and behaviour. We present a neural network implementation (BunDLe Net) and apply it on calcium imaging neuronal data of the roundworm *C. elegans*. Our method reveals clear orbit-like trajectories which are recurrent and structured. It also outperforms conventional methods in the field such as PCA, autoencoders and autoregressors with regards to the dynamical predictability and behavioural decoding accuracy.

Keywords: representation learning; neuronal; cognitive; dynamics; behaviour

Introduction

The rapid development of neuroimaging techniques has resulted in neuronal datasets of ever-increasing detail and complexity. More data, however, does not necessarily translate to a better understanding of brains and neuronal systems. This is because larger datasets, and models of these datasets, are often much harder to interpret (Hoel, Albantakis, Marshall, & Tononi, 2016) even with state-of-the-art tools in computational neuroscience (Jonas & Kording, 2017). One way to deal with this is to create a high-level representation of the neuronal activity (Marr, 1982; Schölkopf et al., 2021).

Commonly-used methods for learning state representations in neuroscience include dimensionality reduction techniques, the majority of which are unsupervised (Kato et al., 2015; Gao et al., 2017). Such approaches have been criticised since they attempt to model the brain in isolation, without recourse to the behaviour it implements (Jonas & Kording, 2017; Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). The resulting state representations are of limited practical use since they are difficult to relate to behaviour, let alone model it.

At the other end of the modelling spectrum are psychologists who create and employ cognitive models to reason about a subject's behaviour. These models are generally arrived at by empirical behavioural studies. This approach is very useful and one often makes causal statements about behaviour and cognitive states, for example: *The boy started crying [behaviour] because he was afraid [cognitive state of fear]*. Such causal statements would become more concrete (and possibly testable) if the cognitive states were to have a grounding in neuronal activity.

In this work we propose a framework to learn cognitive state representations directly from neuronal activity with respect to a behaviour of interest. We first introduce our motivating theoretical principles and a working definition of a cognitive state. Based on this, we propose a generic architecture for learning neuronal state representations from time-series data. We then implement and evaluate our algorithm on neuronal and behavioural data from the nematode *C. elegans*.

Motivating theoretical principles

For the scope of this paper, we propose the following working definition of a cognitive state.

Cognitive state: A high-level representation of neuronal activity that contains sufficient information to model a given set of behaviours and their dynamics.

Let vector X_t represent the neuronal state at time t. We wish to learn a mapping $\tau : X_t \mapsto Y_t$ where Y_t is the desired cognitive state representation at time t. Typically, we want the Y-level to be a lower-dimensional and coarser representation of the X-level. Let T_X and T_Y be transition models at the neuronal and cognitive level respectively.

$$\begin{array}{ccc} X_t & \xrightarrow{\mathbf{T}_X} & X_{t+1} \\ \downarrow \tau & & \downarrow \tau \\ Y_t & \xrightarrow{\mathbf{T}_Y} & Y_{t+1} \end{array}$$

Our working definition requires the cognitive level to preserve dynamical information. This ensures that the cognitive level is self-contained and *sealed-off* from fine-grained details at the neuronal level (Hofstadter, 1979). This can be achieved by requiring the following diagram to commute: i.e. it should not make a difference if we start with X_t and first apply τ and then T_Y or first apply T_X and then τ.

Aside from preserving dynamical information, we require τ to preserve behavioural information so that our abstraction is useful for modelling a specific set of behaviours. At the same time, τ should ideally discard information that is irrelevant to the behaviour so as to keep the representation as succinct as possible. A simple autoencoder framework would be inadequate, since it is based on state-reconstruction and would try to preserve details that are irrelevant to behavioural dynamics (Zhang, McAllister, Calandra, Gal, & Levine, 2021).

Architecture for learning representations

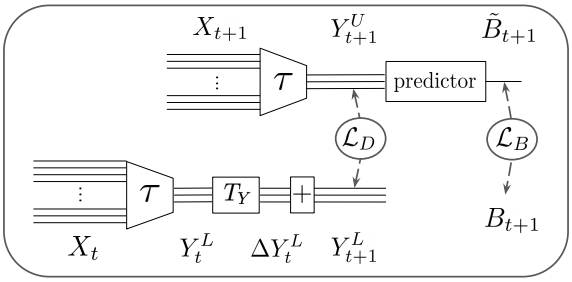


Figure 1: Generic architecture for learning neuronal state representations with behavioural decoding

We present BunDLe Net (**B**ehavio**u**r a**n**d **D**ynamics **Le**arning **Net**work), an algorithm for learning continuous-valued cognitive state representations from neuronal and behavioural time-series data. The architecture is based on the

commutativity diagram, where T_Y and τ are represented by learnable layers[1] in the network. Since we have time-series data, we need not learn T_X, but simply feed in X_{t+1} in the network.

The upper and lower arms correspond to the possible paths in the commutativity diagram. Y^U_{t+1} and Y^L_{t+1} result from the layers in the upper and lower arms respectively. We define a *dynamical loss* $\mathcal{L}_D$ as a mean-squared error that forces Y^L_{t+1} and Y^U_{t+1} to be equalised, thus ensuring the commutativity diagram holds. In the upper channel, Y^U_{t+1} is passed through a predictor layer which outputs $\tilde{B}_{t+1}$ which is trained to match the true behaviour B_{t+1} through a *behavioural loss* $\mathcal{L}_B$ (cross-entropy). This ensures that Y_t contains the same amount of information about B_t as X_t. Both loss terms are weighted by a hyperparameter γ and total loss is given by, $\mathcal{L} = (1-\gamma)\mathcal{L}_B + \gamma\mathcal{L}_D$.

Representation learning on neuronal data

We apply our architecture to learn representations on neuronal data from the nematode *C. elegans* which consists of time-series recording of 109 neurons (Kato et al., 2015). The behavioural data is a time-series of human-annotated behaviours that denote the motor state of the worm. The dimensionality of the latent space was chosen to be 3 for ease of visualisation.

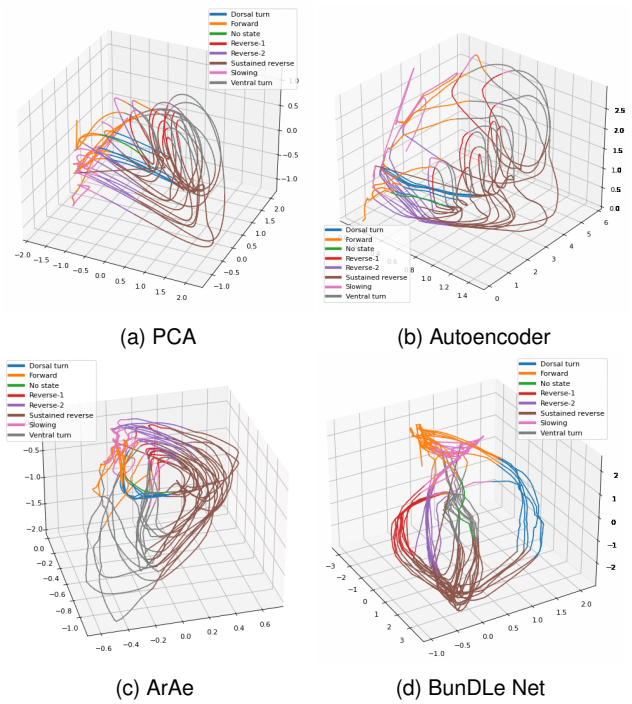
To evaluate our algorithm, we compare it with typically-used representation learning methods in neuroscience, such as PCA, autoencoder, and an autoregressor implemented with an autoencoder architecture (ArAe). The autoencoder was trained with a standard reconstruction loss, while the ArAe attempts to reconstruct X_{t+1} from X_t. In Figure 2, we observe that all representations capture the *recurrent* nature of the dynamics. For both PCA and the autoencoder however, there is a drift that drags out the dynamics in an arbitrary direction thus mapping every sample to a different point in state space. In the representations from ArAe and BunDLe Net, coarse-graining occurs in a truer sense i.e. neuronal-level information irrelevant to the behaviour is discarded.

In BunDLe Net's embedding, we observe branching orbit-like trajectories. Note that the learned dynamics is largely deterministic along a give branch. It is only at the bifurcation points where stochasticity is seen to emerge. These bifurcation points may be interpreted as where probabilistic decisions at the cognitive level are made. Thus, the algorithm distils out the stochasticity and confines it to local regions in state space. This is a direct result of requiring dynamical information (that is behaviourally relevant) to be preserved. Owing to this, the learned representation is visually interpretable and reveals a structured nature of the cognitive state space.

Discussion

In this work, we have presented a generic architecture to learn cognitive state representations from neuronal data based on a few simple but vital principles. It is noteworthy that our architecture is similar to some of those used in the field of contrastive learning on image data (Kipf, van der Pol, & Welling, 2020; Illing, Ventura, Bellec, & Gerstner, 2021) but with an added behavioural context. Having demonstrated BunDLe Net on *C. elegans* data we see that it outperforms other standard algorithms. This work could be useful in gaining automated insights from neuroscience data. There is also potential in extending this work in the direction of decision making based on the bifurcation points in the learned representations. Finally, this work is a small step in creating causal cognitive models with a neuronal basis. Further work would involve developing a causal framework for such an algorithm.

[1]These layers are not restricted to typical ANN layers but can be implemented by any transition model including variational layers



(a) PCA (b) Autoencoder

(c) ArAe (d) BunDLe Net

Figure 2: Dynamics of the neuronal data in the 3-dimensional latent space learned by various algorithms

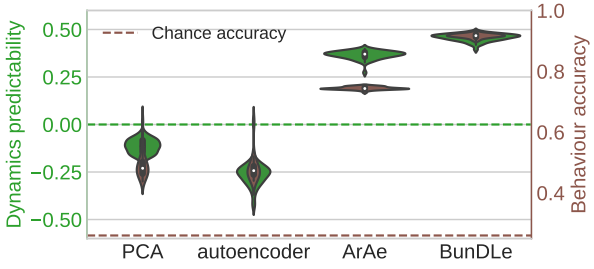


Figure 3: Quantitative evaluation based on behavioural decoding accuracy (right-axis, brown) and dynamical predictability (left-axis, green) which is measured by MSE between predicted and true Y_{t+1} normalised with respect to a baseline copy-input-to-output autoregressor

References

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., & Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*. doi: 10.1101/214262

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016, 08). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1). (niw012)

Hofstadter, D. R. (1979). Gödel, escher, bach: An eternal golden braid. In (20th-anniversary Edition ed.). Basic Books Inc.

Illing, B., Ventura, J., Bellec, G., & Gerstner, W. (2021). Local plasticity rules can learn deep representations using self-supervised contrastive predictions. In *Advances in neural information processing systems* (Vol. 34, pp. 30365–30379). Curran Associates, Inc.

Jonas, E., & Kording, K. P. (2017, January). Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, *13*(1), e1005268.

Kato, S., Kaplan, H. S., Schrödel, T., Skora, S., Lindsay, T. H., Yemini, E., ... Zimmer, M. (2015, October). Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, *163*(3), 656–669.

Kipf, T. N., van der Pol, E., & Welling, M. (2020). Contrastive learning of structured world models. In *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, april 26-30, 2020*. OpenReview.net.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480-490.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., & Levine, S. (2021). Learning invariant representations for reinforcement learning without reconstruction. In *9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021*. OpenReview.net.