Continual learning in artificial neural networks as a computational framework for understanding representational drift in neuroscience

Daniel Anthes (danthes@uni-osnabrueck.de)
Institute of Cognitive Science, Wachsbleiche 27
49090 Osnabrück, Germany

Sushrut Thorat (sthorat@uni-osnabrueck.de)
Institute of Cognitive Science, Wachsbleiche 27
49090 Osnabrück, Germany

Peter König* **(pkoenig@uni-osnabrueck.de)**
Institute of Cognitive Science, Wachsbleiche 27
49090 Osnabrück, Germany

Tim C Kietzmann* **(tkietzma@uni-osnabrueck.de)**
Institute of Cognitive Science, Wachsbleiche 27
49090 Osnabrück, Germany
* shared last author

Abstract

Studies monitoring neural responses over time have shown that neural representations "drift", while behaviour stays constant - a phenomenon suggested to be linked to learning. Here we demonstrate that continual learning in deep neural networks may serve as a modelling framework for making progress in this domain, (a) for understanding the underlying computations and (b) for testing the analysis tools used. We train networks that implement two different neuroscientific theories on how stable behaviour can be maintained in light of learning new tasks. The first strategy allows for the models' readouts to 'track' the changing representations. The second confines learning to the nullspaces of previously learned readouts. Both simulations replicate hallmarks of drift observed in neuroscience - changing single-unit tuning, reduced cross-decoding performance over time, and changes in the overall population response. At the same time, existing analysis techniques cannot reliably differentiate the two implemented mechanisms. Continual learning may therefore offer a language for expressing computational hypotheses on drift, as well as a testbed for developing new analysis techniques.

Keywords: Representational Drift; Continual Learning; Artificial Neural Networks; Normative Models

Introduction

Understanding the neural representations underlying behaviour has been a longstanding enterprise in neuroscience. Curiously, neural recordings performed over multiple weeks have shown that representations "drift" while behaviour stays constant (Driscoll, Pettit, Minderer, Chettih, & Harvey, 2017). Whether drift is a bug or a feature remains a topic of debate (Masset, Qin, & Zavatone-Veth, 2022), but one emerging view suggests that drift may be the result of, or even contribute to, learning (Driscoll, Duncker, & Harvey, 2022; Micou & O'Leary, 2023). In line with this, work in AI continual learning (CL) has focused on how representations change as new tasks are trained (Anthes, Thorat, König, & Kietzmann, 2023; Davari, Asadi, Mudur, Aljundi, & Belilovsky, 2022), suggesting that "drift" may support a system's ability to keep learning, while maintaining performance on previous tasks (Anthes, Thorat, König, & Kietzmann, 2024). Irrespective of its role, drift raises the question about how the brain "reads out" stable information, while the underlying representations change (Rule et al., 2020). One suggestion is that the readout is updated along with the representations, e.g. via Hebbian plasticity (Rule & O'Leary, 2022). Another strategy discussed is that no updating is needed, as drift occurs in an orthogonal subspace (Rule, O'Leary, & Harvey, 2019).

In this work we present normative models for both strategies. This is accomplished by adapting techniques from continual learning in artificial intelligence. We first verify that hallmarks of drift occur in these models, and then draw conclusions about the ability of commonly used analysis techniques to differentiate such computational strategies to compensate for changing representations.

Methods

We consider two CL algorithms as normative models implementing two hypotheses from computational neuroscience that describe how stable behaviour can be accomplished despite drifting representations.

Strategy A This strategy is based on the Learning without Forgetting (LwF) algorithm (Li & Hoiem, 2017), whose gradient-based procedure is similar to the Hebbian plasticity procedure from previous work (Rule & O'Leary, 2022) modelling biological drift. Before learning a new task, pseudo-labels for the new task images are generated at the old tasks' readouts. While learning the new tasks, the old readouts are trained to maintain a stable mapping between the new task's inputs to their pseudo-labels, effectively maintaining stability for the old task mappings while allowing learning on the new task. In this setting, the old readouts change with the drifting representations in the network (Fig. 1A, top).

Strategy B This strategy restricts updates to previously learned representations to the nullspace of old readouts (Saha, Garg, & Roy, 2021; Farajtabar, Azizan, Mott, & Li, 2020). While learning a new task, gradients for previous tasks with respect to the activations are computed at each layer of the network. Gradients for the new task are projected into the nullspace of activation gradients for the old task. This ensures that activations for the previous tasks do not change in ways that would affect their readouts. In this setting, the old readouts do not adapt and representations continue to drift in the nullspace of those readouts (Fig. 1A, bottom).

Simulation setup

We sequentially train a fully connected ANN (two hidden layers with 256 units, ReLU activations) on six unique subsets sampled from the combination of MNIST (LeCun, Bottou, Bengio, & Haffner, 1998), Kuzushiji-MNIST (Clanuwat et al., 2018), and Fashion-MNIST (Xiao, Rasul, & Vollgraf, 2017). Each subset constitutes a 5 way-classification task, for which a separate readout is trained. We trained 10 instances for each strategy. In all the results, we report the mean and 95% confidence intervals thereof across these instances.

Results

Models employing either strategy showed stable behaviour (classification accuracy) on the first task as they learned new tasks (Fig. 1C, black line), as compared to the baseline where no continual learning algorithm was deployed (dotted gray line).

Representational Drift Analyses

Here, we assess how representations in the computational models change as they learn new tasks - to do so, we adapt well-known techniques from the neuroscientific literature to check for signatures of drift.

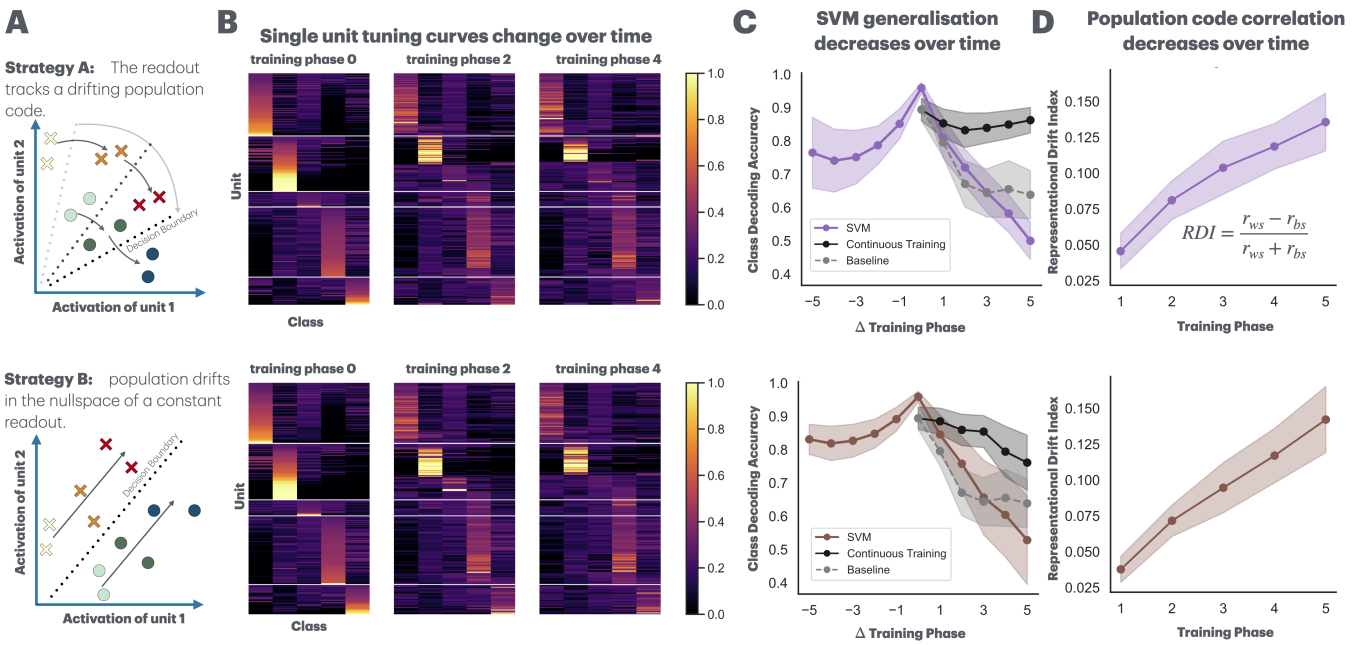

Figure 1: **A:** Illustration of different scenarios for the effects of representational drift on downstream areas. **B:** Single neuron tuning curves. Each row denotes the distribution of activity over all 5 classes in the first task (normalised to sum to 1 for each phase). **C:** cross decoding performance for SVMs trained on representations at different training phases. **D:** population code correlation measured with the representational drift index introduced in Schoonover et al. (2021). r_{ws} denotes the correlation between random halves of population responses from the same phase, r_{bs} denotes correlation between random halves of responses from different phases (Computed for each class included in the first task. Shown is the mean over classes).

First, considering single neuron tuning, prior studies have shown that single neurons in mice change their tuning over days/weeks while the animal maintains stable behaviour (Driscoll et al., 2017). Analogously, we checked if tuning to classes across units in the pre-readout layer of the models changed similarly over training phases. As seen in Fig. 1B, both models showed tuning changes while learning. Intriguingly, the most selective units had the largest changes in their class tuning.

Second, considering task-relevant subspaces in the population code, a prior study showed that linear SVMs trained on representations on a given day do not generalise to representations on another day/week (Schoonover et al., 2021). Generalisation decreases with increasing time intervals between the train and test days. As seen in Fig. 1C, our two models both show this pattern - SVMs trained on a given training phase, for task 1 data, generalise poorly to other training phases, with the generalisation depending on the temporal interval between those phases. For Strategy B, this loss of generalisation is surprising, as we know that a stable readout exists, as evidenced by the black line in Fig. 1C (bottom panel). This can be explained by the usage of different readout strategies: while the models use linear softmax classifiers, the analysis relies on SVMs. These results suggest that these SVMs do not rely on the same subspace as the natural readout of the continual learners. This finding has implications for analyses of neural drift, as the brain's readout strategy remains unknown - choosing the wrong readout analysis strategy may therefore overestimate the consequences of drift.

Third, assessing overall representational changes, Marks and Goard (2021) have shown that the correlations of stimulus representations across days decrease with increasing time intervals between those days. As seen in Fig. 1D, and assessed via representational drift index (RDI), both of our models show this pattern of an increasing RDI, i.e. class representational correlations (from task 1) across training phases decrease with increasing temporal intervals between those phases.

Conclusion

The role of representation drift in the brain remains an open question. Here, we adhere to the hypothesis that drift is essential for learning and presented two computational models that express the two main hypotheses of how the brain may deal with drifting representations: adapting the readout, or confining drift to the readout's nullspace. We demonstrate that both systems exhibit drift, as measured via traditional drift analysis techniques from computational neuroscience. However, none of the analyses could reliably differentiate between these two accounts. These findings suggest that continual learning in deep neural networks could be used as a framework for implementing computational hypotheses on drift (Doerig et al., 2023; Golan et al., 2023), and for the development of new analysis techniques in neuroscience, thereby further elucidating the nature of drift in biological systems.

Acknowledgments

The project was financed by the funds of the research training group “Computational Cognition” (GRK2340) provided by the Deutsche Forschungsgemeinschaft (DFG), Germany and the European Union (ERC, TIME, Project 101039524). Compute resources used for this project are additionally supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 456666331.

References

Anthes, D., Thorat, S., König, P., & Kietzmann, T. C. (2023). Diagnosing catastrophe: Large parts of accuracy loss in continual learning can be accounted for by readout misalignment.

Anthes, D., Thorat, S., König, P., & Kietzmann, T. C. (2024). *Keep moving: identifying task-relevant subspaces to maximise plasticity for newly learned tasks.*

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*.

Davari, M., Asadi, N., Mudur, S., Aljundi, R., & Belilovsky, E. (2022). Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16712–16721).

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., ... others (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Driscoll, L. N., Duncker, L., & Harvey, C. D. (2022). Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, *76*, 102609.

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., & Harvey, C. D. (2017). Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, *170*(5), 986–999.

Farajtabar, M., Azizan, N., Mott, A., & Li, A. (2020). Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics* (pp. 3762–3773).

Golan, T., Taylor, J., Schütt, H., Peters, B., Sommers, R. P., Seeliger, K., ... others (2023). Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *Behavioral and Brain Sciences*, *46*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935–2947.

Marks, T. D., & Goard, M. J. (2021). Stimulus-dependent representational drift in primary visual cortex. *Nature communications*, *12*(1), 5169.

Masset, P., Qin, S., & Zavatone-Veth, J. A. (2022). Drifting neuronal representations: Bug or feature? *Biological cybernetics*, *116*(3), 253–266.

Micou, C., & O'Leary, T. (2023). Representational drift as a window into neural and behavioural plasticity. *Current opinion in neurobiology*, *81*, 102746.

Rule, M. E., Loback, A. R., Raman, D. V., Driscoll, L. N., Harvey, C. D., & O'Leary, T. (2020). Stable task information from an unstable neural population. *elife*, *9*, e51121.

Rule, M. E., & O'Leary, T. (2022). Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences*, *119*(7), e2106692119.

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current opinion in neurobiology*, *58*, 141–147.

Saha, G., Garg, I., & Roy, K. (2021). Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*.

Schoonover, C. E., Ohashi, S. N., Axel, R., & Fink, A. J. (2021). Representational drift in primary olfactory cortex. *Nature*, *594*(7864), 541–546.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.