Diagnosing Catastrophe: Large Parts of Accuracy Loss in Continual Learning Can Be Accounted for by Readout Misalignment

Daniel Anthes (danthes@uos.de)
Institute of Cognitive Science, Osnabrück University
Wachsbleiche 27, 49090 Osnabrück, Germany

Sushrut Thorat (sushrut.thorat94@gmail.com)
Institute of Cognitive Science, Osnabrück University
Wachsbleiche 27, 49090 Osnabrück, Germany

Peter König[*] (pkoenig@uos.de)
Institute of Cognitive Science, Osnabrück University
Wachsbleiche 27, 49090 Osnabrück, Germany

Tim C Kietzmann[*] (tkietzma@uos.de)
Institute of Cognitive Science, Osnabrück University
Wachsbleiche 27, 49090 Osnabrück, Germany
[*] shared last author

Abstract

Unlike primates, training artificial neural networks (ANNs) on changing data distributions leads to a rapid decrease in performance on old tasks. This phenomenon is commonly referred to as catastrophic forgetting. In this paper, we investigate the representational changes that underlie this performance decrease and identify three distinct processes that together account for the phenomenon. The largest component is a misalignment between hidden representations and readout layers. Misalignment occurs due to learning on additional tasks and causes internal representations to shift. Representational geometry is partially conserved under this misalignment and only a small part of the information is irrecoverably lost. All types of representational changes scale with the dimensionality of hidden representations. These insights have implications for deep learning applications that need to be continuously updated, but may also aid aligning ANN models to the rather robust biological vision.

Keywords: continual learning; catastrophic forgetting; artificial neural networks; representations; lifelong learning

Introduction

Our world is inherently sequential. Adapted to this, humans are successful in continuously learning new skills over their lifetime. However, most state-of-the-art training procedures for artificial neural networks (ANNs) rely on data being independent and identically distributed. In settings where the data distribution changes, networks have been reported to rapidly forget previous knowledge (Parisi, Kemker, Part, Kanan, & Wermter, 2019; Hadsell, Rao, Rusu, & Pascanu, 2020). This phenomenon is commonly termed *catastrophic forgetting* (French, 1999; McCloskey & Cohen, 1989).

A number of factors influence the degree to which performance decreases in sequential learning scenarios: the dimensionality of representations (Mirzadeh et al., 2022), pretraining (Ramasesh, Lewkowycz, & Dyer, 2022), objective function (S. Li, Du, van de Ven, & Mordatch, 2022; Davari, Asadi, Mudur, Aljundi, & Belilovsky, 2022) and task similarity (Ramasesh, Dyer, & Raghu, 2020). However, the changes to the task-relevant representations during continual learning remain to be fully characterized (see Davari et al. (2022) for first steps). In this work, we characterize changes in representational geometry and their contribution to the observed decrease in performance. We find that rather than forgetting, much of the degraded performance can be explained by a misalignment of representations and the readouts of the network.

Analysis

Our model system is a standard four layer convolutional network The training procedure, task and network architecture are identical to Zenke, Poole, and Ganguli (2017). We study catastrophic forgetting in the task-incremental scenario (Van de Ven & Tolias, 2019), initializing a new classification head every time a novel task is encountered. After pretraining the network on CIFAR10 (Krizhevsky & Hinton, 2009), we sequentially train on ten equal task splits from CIFAR-100. We repeat this procedure 5 times controlling for the effects of task similarity by randomly assigning each class to a task (Ramasesh et al., 2020).

We characterise the information present throughout learning by training diagnostic readouts for all tasks after every phase of training. A drop in performance, despite adjusted readout, constitutes a loss of task relevant information. This scenario constitutes true *forgetting*. Contrary to this, performance loss attributed to *misalignment* is computed by the difference in performance between the original readout ($t = 0$) and the newly trained diagnostic readouts at every phase of training. Third, to estimate the extent to which misalignment is due to rotation, translation, and uniform scaling of an otherwise static geometry, we align representations for each task after each training phase to the representations immediately after learning the task ($t = 0$) with a geometry-preserving Procrustes transformation (Gower, 1975).

Finally, as increasing layer width has been shown to alleviate catastrophic forgetting (Mirzadeh et al., 2022), we vary the width of the final hidden layer to investigate how the different components of representational change are modulated by network capacity.

Results

As expected, we observe effects of catastrophic forgetting, i.e. a rapid drop in performance of the original readouts as the network is trained on additional tasks (Fig. 1A, 'continual' at T$>$0). Notably, however, performance of diagnostic readouts decreases much less, indicating that the discriminability of the old classes is indeed preserved, i.e. there is little "actual" forgetting. The primary cause of decreased performance is readout misalignment, the extent of which is shown by the large difference between 'continual' performance and performance measured at the diagnostic readouts (Fig. 1A, 'diagnostic'), in line with similar previous analyses (Davari et al., 2022).

Does misalignment preserve the original representational geometry? If so, we'd expect that Procrustes alignment should yield performance as good as the linear diagnostic readouts. We observe that aligning representations accounts for approximately half of the performance difference between continual and diagnostic readouts (Fig. 1 A, 'procrustes'). Therefore, misalignment can be characterized as a combination of geometry preserving and deforming changes of representations.

An open question that remains from our and previous work is whether the comparably good performance of the diagnostic readout is explained by transfer learning based on features learned for earlier tasks in the sequence. Indeed, we observe that the features learned for previously encountered tasks transfer to unseen tasks ('Feature Transfer' in Fig. 1). Yet, transfer cannot fully explain the performance observed with diagnostic readouts, as a clear discontinuity in the diagnostic readout performance trajectory from before to after training a new task ($t = 0$) can be seen. This suggests that

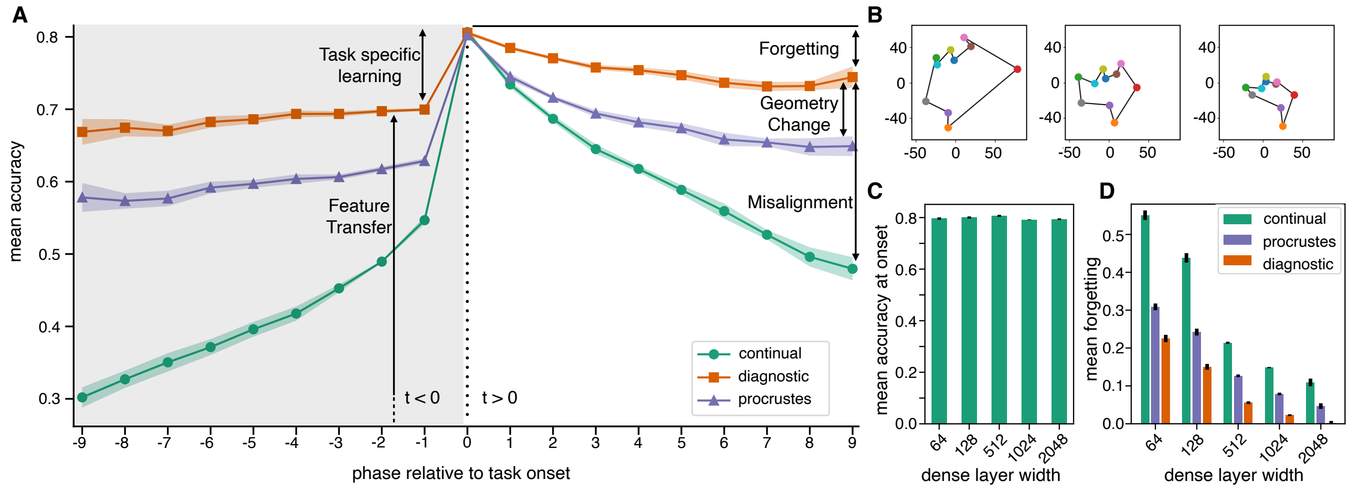

Figure 1: **A:** Classification accuracy averaged over the ten tasks sampled from CIFAR100. Prior to averaging, task performance trajectories are temporally aligned to task onset such that the x-axis reflects performance after t additional tasks have been learned. The shaded area around each line indicates the standard error computed over five repetitions of the procedure. **B:** Mean class representations for a task split. Task mean representations from all phases are projected to a shared two dimensional space using multidimensional scaling (Torgerson, 1952). Shown are representation vectors directly after learning the task, after learning 5, and after learning 9 additional tasks (left to right). **C:** Mean performance over all tasks at task onset ($t = 0$). Network size does not have an effect on how well tasks are learned initially. **D:** Performance loss measured as the difference between performance at $t = 0$ and the mean over performances measured at all $t > 0$. Standard error for additional networks is computed over three simulations.

newly learned features better support the new task. This additional information stays preserved in the network over learning of multiple additional tasks, as evidenced by the fact that diagnostic readout performance stays above the performance measured at $t = -1$ for the subsequent phases ($t > 0$).

Finally, characterizing the influence of network size in continual learning with our new analysis techniques, we find that varying the width of the final hidden layer attenuates all three measures of representational change. Yet, we still observe small amounts of changes to the representational geometry and misalignment with the readouts of the respective networks (Fig. 1 C & D).

Discussion

In characterizing representational changes in a neural network during continual learning, we observed that misalignment of the pre-readout representations with the task readouts explains large parts of performance degradation that is commonly referred to as ‘catastrophic forgetting’. Interestingly, only a small amount of performance cannot be linearly read out and is irrecoverably ‘forgotten’.

Many algorithms addressing catastrophic forgetting rely on restricting learning at synapses that encode information for previous tasks (Zenke et al., 2017; Kirkpatrick et al., 2017) or regularize learning of representations for new tasks (Z. Li & Hoiem, 2017) in order to not lose information relevant for the previous tasks. We argue that information in hidden layers is largely preserved, even without restricting learning trajectories or placing constraints on representations the network is allowed to learn. This is especially prominent in larger networks. We hypothesize that catastrophic forgetting may instead be efficiently addressed by solving the problem of readout misalignment without influencing the learning of new tasks (See also: (Lesort, George, & Rish, 2021)). Indeed, there may be benefits to not restricting learning of representations more than necessary, as restrictions to the learning dynamics of the network may lead to decreased plasticity or sub-optimal solutions over long sequences of tasks.

Lastly, the primate visual system is successfully able to learn new tasks without exhibiting forgetting of old tasks. If we are to use ANNs as models of biological vision, then the discrepancies in the learning dynamics of the two systems remain to be addressed. Future work will test the currently described analysis framework for characterising representational changes in continual learning on biological data to further understand where, how, and when the visual system copes with the newly arriving information.

Acknowledgments

The project was financed by the Deutsche Forschungsgemeinschaft (DFG, research training group “Computational Cognition”, GRK2340), as well as the European Research Council (ERC, TIME, Project 101039524). Compute resources used for this project are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 456666331.

References

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).

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*, 33–51.

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, *24*(12), 1028–1040.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Lesort, T., George, T., & Rish, I. (2021). Continual learning in deep networks: an analysis of the last layer. *arXiv preprint arXiv:2106.01834*.

Li, S., Du, Y., van de Ven, G., & Mordatch, I. (2022). Energy-based models for continual learning. In *Conference on lifelong learning agents* (pp. 1–22).

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

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Mirzadeh, S. I., Chaudhry, A., Yin, D., Hu, H., Pascanu, R., Gorur, D., & Farajtabar, M. (2022). Wide neural networks forget less catastrophically. In *International conference on machine learning* (pp. 15699–15717).

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, *113*, 54–71.

Ramasesh, V. V., Dyer, E., & Raghu, M. (2020). Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*.

Ramasesh, V. V., Lewkowycz, A., & Dyer, E. (2022). Effect of scale on catastrophic forgetting in neural networks. In *International conference on learning representations.*

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, *17*(4), 401–419.

Van de Ven, G. M., & Tolias, A. S. (2019). Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. In *International conference on machine learning* (pp. 3987–3995).