MANCHESTER 1824
The University of Manchester

The University of Manchester Research

One system for learning and remembering episodes and rules

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Hewson, J. T. S., Sloman, S. J., & Dubova, M. (2024). *One system for learning and remembering episodes and rules*.

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:15. Jul. 2024

One system for learning and remembering episodes and rules

Joshua T. S. Hewson (joshua_hewson@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI

Sabina J. Sloman (sabina.sloman@manchester.ac.uk)
Department of Computer Science, University of Manchester, Manchester, UK

Marina Dubova (mdubova@iu.edu)
Cognitive Science Program, Indiana University, Bloomington, IN

arXiv:2407.05884v1 [cs.AI] 8 Jul 2024

Abstract

Humans can learn individual episodes and generalizable rules and also successfully retain both kinds of acquired knowledge over time. In the cognitive science literature, (1) learning individual episodes and rules and (2) learning and remembering are often both conceptualized as competing processes that necessitate separate, complementary learning systems. Inspired by recent research in statistical learning, we challenge these trade-offs, hypothesizing that they arise from capacity limitations rather than from the inherent incompatibility of the underlying cognitive processes. Using an associative learning task, we show that one system with excess representational capacity can learn and remember both episodes and rules.

Keywords: Remembering, catastrophic forgetting, complementary learning systems, continual learning

Introduction

In the study of learning, two trade-offs have historically been observed in the behavior of computational models: (1) between the abilities to simultaneously learn individual episodes and generalizable rules and (2) between the abilities to learn and to remember. For example, connectionist models often show that (1) memorizing individual episodes leads to a reduced ability to learn the rules required to generalize to new episodes ("overfitting") (McClelland, McNaughton, & O'Reilly, 1995) and (2) learning in a new task leads to catastrophic forgetting of what has been learned in previous tasks (McCloskey & Cohen, 1989). These observations motivated the creation of dual-system theories, such as the complementary learning systems model (McClelland et al., 1995), which posit separate learning systems for learning and remembering episodes and rules.

Recent research has shown that the trade-off between learning episodes and rules is not inherent to learning in computational systems. The computational models in which these trade-offs were historically observed had limited *capacity*: They could memorize only a small number of their observations. Computational systems with excess capacity – which can recover far more relationships between the features of observations – have the ability to both memorize and generalize, i.e., to learn both episodes and rules (Dubova & Sloman, 2023; Belkin, Hsu, Ma, & Mandal, 2019; Nakkiran et al., 2019; Davies, Langosco, & Krueger, 2023). In this study, we demonstrate that excess capacity systems can also overcome the apparent trade-off between learning and remembering, i.e., they can simultaneously successfully learn new episodes and rules *and* remember previously-learned episodes and rules.

Methods

Catastrophic forgetting. Human participants in the behavioral test referenced by McClelland et al. (1995) were tasked with memorizing batches of random word pairings in a blocked regime (Barnes & Underwood, 1959). During the first block, participants were presented with a list of words (list A) and tasked with memorizing arbitrary associations between the words on list A and the words on another list B ($A-B$ pairings). During the second block, they were presented with a new word list C and tasked with memorizing arbitrary associations between the words on list A and on list C ($A-C$ pairings). Over the course of training on the $A-C$ pairings, participants were tested on the $A-B$ pairings they learned during the first block. Participants showed memory interference, but were still able to retain most of the previously learned associations. McCloskey and Cohen (1989) modeled behavior in this task with a simple connectionist model. This model forgot nearly all information about the $A-B$ pairings after being trained on the $A-C$ pairings, a phenomenon they referred to as *catastrophic forgetting*.

Model. Inspired by McCloskey and Cohen (1989), we used a simple multi-layer perceptron architecture with two hidden layers of equal width.[1]

Task. We expand on McCloskey and Cohen (1989)'s procedure by changing the data to vary on a continuum from rules to episodes, so that the dynamics of learning and forgetting of arbitrary associations between episodes and generalizable rules can be studied together.

Two sample datasets of 10 5-dimensional samples, A_{train} and A_{test}, were created by sampling from a Gaussian probability distribution. These datasets were then passed through a transformation f. Two target datasets, B and C, were each formed by taking a weighted sum between the transformed sample data and a set of random perturbations. A third target dataset D was created by removing the random perturbations from C:

$$
\begin{gathered}
A_{train} \sim \mathcal{N}(0,1) \\
A_{test} \sim \mathcal{N}(0,1) \\
B = (1-noise)\cdot f(A_{train}) + noise \cdot \varepsilon_B \\
C = (1-noise)\cdot f(A_{test}) + noise \cdot \varepsilon_C \\
D = (1-noise)\cdot f(A_{test})
\end{gathered}
$$

where $0 \leq noise \leq 1$, $\varepsilon_B \sim \mathcal{N}(0,1)$ and $\varepsilon_C \sim \mathcal{N}(0,1)$. f represents the generalizable rule that characterizes the relationship between the sample and corresponding target data (i.e., that characterizes each of the $A_{train}-B$, $A_{test}-C$ and $A_{test}-D$ pairings). The $noise$ parameter controls the amount of structure in the data: When $noise = 0$, the task amounts entirely to learning of the generalizable rule; when $noise = 1$, the task amounts entirely to learning arbitrary associations between the sample data and episodes ε_B (in the $A_{train}-B$ pairings) and ε_C (in the $A_{test}-C$ pairings).

Capacity. Our key manipulation was the *capacity* of each model we tested. The capacity of a model is defined as the minimum number of hidden nodes needed to fully memorize a given dataset. *Constrained capacity* models have fewer hidden nodes than required to memorize the data they are

[1] Find our code at: `https://github.com/TheLemonPig/ECLvsCLS`

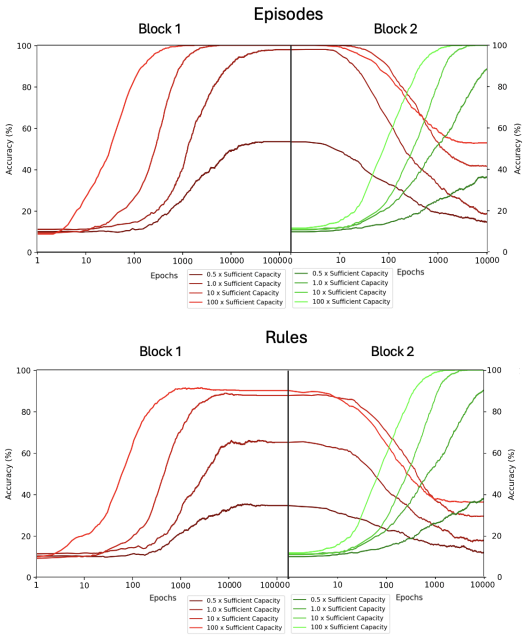

Figure 1: Temporal plots for mean classification accuracy over training (the noise level is fixed at 25%). Left: The episode (top) or rule (bottom) for $A_{train} - B$ is learned. Right: The episode $A_{train} - C$ is learned (green lines) while the episode (top) or rule (bottom) for $A_{train} - B$ is being forgotten.

presented with. *Sufficient capacity* models have just enough nodes to memorize the data they are presented with. *Excess capacity* models have more nodes than required to memorize the data they are presented with.

We tested models with capacities of .5, 1, 10 and 100 times the capacity needed to fully memorize the datasets (constrained, sufficient, excess and excess capacity, respectively).

Training. During Block 1, the models were trained to associate A_{train} with B, which involves learning both the rule f and the arbitrary component of the episodes, ε_B. During Block 1, we also tested the models' abilities to generalize from A_{test} to C. During Block 2, the models were trained to associate A_{train} with C. During Block 2, we also tested the models' abilities to recall the $A_{train} - B$ pairings and to predict the $A_{test} - D$ pairings, which capture the models' abilities to remember episodes and rules, respectively (we test learning of the rule on the basis of performance on the $A_{test} - D$ pairings in order to isolate error from failure to learn f from error caused by the noise added to C).

The models were optimized with Stochastic Gradient Descent using a mean squared error loss function (learning rate = 0.01). All models were trained until convergence, defined as a rate of decrease in loss going below 1×10^{-5} per $5,000$ epochs. We ran all simulations 100 times.

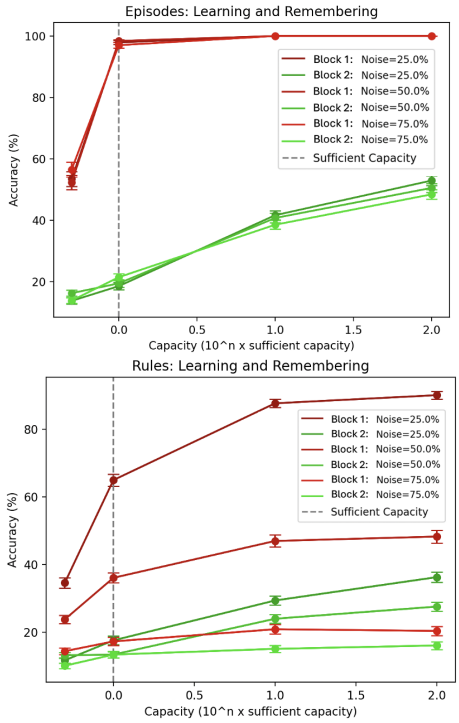

Figure 2: Final averaged mean results after training on Block 1 and 2 respectively, with varying levels of noise. Left of the dashed line: constrained capacity; Dashed line: sufficient capacity, Right of the dashed line: excess capacity. Error bars show standard errors.

Results

Consistent with prior literature (e.g., Belkin et al. (2019); Nakkiran et al. (2019)), in all cases, the models with excess capacity were better able than models with constrained and sufficient capacity to learn both episodes and rules (Fig. 2). The difference between the mean classification accuracy of all models was statistically significant ($p < 0.001$) for both learning episodes, defined by the t-test of the means between each pair of models, at the end of training. In other words, there was not a consistent trade-off between learning episodes and learning rules in the excess capacity regime.

Consistent with prior work on catastrophic forgetting, models with constrained and sufficient capacity exhibited only a limited ability to retain prior knowledge when having to learn a new set of interfering associations. The models with excess capacity were better able to retain knowledge of both episodes and rules (Figs. 1 and 2). All pairwise comparisons between the means of performance of excess vs. constrained/sufficient capacity models were significant at the $p < .001$ level.

Conclusion

Our results demonstrate the in-principle ability of one computational learning system to both learn and remember episodes

and rules. By challenging the traditional view of learning and remembering episodes and rules as inherently competing processes, this work opens new avenues for understanding the flexibility and nuance of cognitive function by exploring the properties of learning in different capacity regimes. Our findings also have important implications for the study of continual learning, transfer learning, and the development of more advanced cognitive architectures (Mannering & Jones, 2021; van de Ven, Soures, & Kudithipudi, 2024; Achille, Rovere, & Soatto, 2019; Sherman, Turk-Browne, & Goldfarb, 2023; Schapiro, Turk-Browne, Botvinick, & Norman, 2017; Liu et al., 2022).

References

Achille, A., Rovere, M., & Soatto, S. (2019). *Critical learning periods in deep neural networks.*

Barnes, J. M., & Underwood, B. J. (1959). "fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*(2), 97–105. Retrieved from https://doi.org/10.1037/h0047507 doi: 10.1037/h0047507

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Davies, X., Langosco, L., & Krueger, D. (2023). *Unifying grokking and double descent.*

Dubova, M., & Sloman, S. J. (2023). Excess capacity learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45). Retrieved from https://escholarship.org/uc/item/49w48008

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., & Williams, M. (2022). Towards understanding grokking: An effective theory of representation learning. In *Advances in neural information processing systems* (Vol. 35, pp. 34651–34663).

Mannering, W. M., & Jones, M. N. (2021). Catastrophic interference in predictive neural network models of distributional semantics. *Computational Brain & Behavior*, *4*, 18–33. Retrieved from https://doi.org/10.1007/s42113-020-00089-5 doi: 10.1007/s42113-020-00089-5

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995, July). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. doi: 10.1037/0033-295X.102.3.419

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Academic Press. Retrieved from https://doi.org/10.1016/S0079-7421(08)60536-8 doi: 10.1016/S0079-7421(08)60536-8

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). *Deep double descent: Where bigger models and more data hurt.*

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049.

Sherman, B. E., Turk-Browne, N. B., & Goldfarb, E. V. (2023). Multiple memory subsystems: Reconsidering memory in the mind and brain. *Perspectives on Psychological Science*. doi: 10.1177/17456916231179146

van de Ven, G. M., Soures, N., & Kudithipudi, D. (2024). *Continual learning and catastrophic forgetting.*