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

Abstract

Humans can learn individual episodes and generalizable rules and also successfully retain 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 episodes and generalizable rules and (2) between the abilities to learn and to remember. For example, connectionist models exhibit the behaviors that (1) memorizing individual episodes leads to a reduced ability to learn the rules required to generalize to new episodes ("overfitting") 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, McNaughton, & O'Reilly, 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 and more complex 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.*

Task. 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. This ratio of rule and episode is controlled by a noise parameter, which at max created a rule-free episode, and at zero created a simple rule.

Data. Two sample datasets of 10 5-dimensional samples, A_{train}, A_{test} are created by sampling from a Gaussian probability distribution. These datasets are then passed through a transformation f. Two target datasets, B and C, are each formed by taking a weighted sum between the transformed data and another sample dataset from the same Gaussian distribution. A third dataset D is created by omitting the added noise to dataset 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 $0 \leq noise \leq 1$. We test this generalization using $A_{test} - D$ because this allows us to ignore the error caused by the noise added to C.

Model. Following McCloskey and Cohen (1989), we used a simple multi-layer perceptron architecture with two hidden layers of equal width. 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. The width of the models' layers varied proportionally with the sufficient capacity relative to each training dataset. We tested models with a capacity of .5, 1, 10 and 100 times the capacity needed to fully memorize the datasets.

Training. During Phase 1, the models were trained to associate A_{train} with B, which involves learning both the rule f and the episodes ε_B. During Phase 1, we also tested the models' abilities to generalize A_test to C. During Phase 2, the models

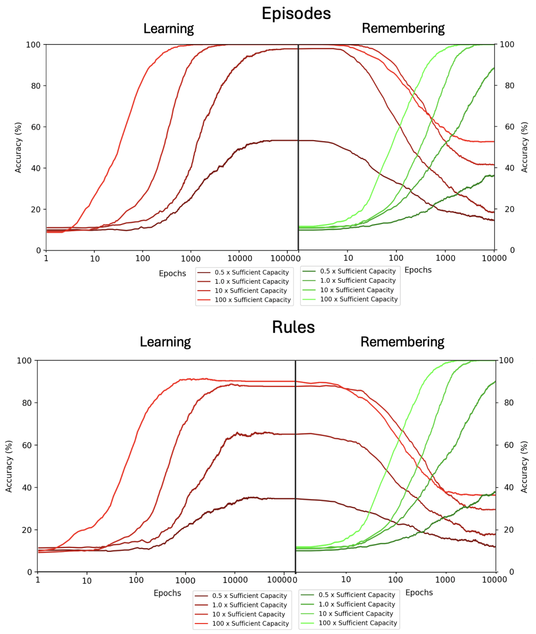

Figure 1: Temporal plots for learning and remembering (the noise level is fixed at 25%). Left: The episode (top) or rule (bottom) for $A_{train} - B$ is learned. Right: The episode $A_{train} - C$ is learned (green lines) while the episode (top) or rule (bottom) for $A_{train} - B$ is being forgotten.

were trained to associate A_{train} with C. During Phase 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.

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.

Results

Consistent with prior literature (e.g. (Belkin et al., 2019; Nakkiran et al., 2019)), the systems with excess capacity overcame the rules vs. episodes trade-off showing superior ability to acquire both episodes ($t = 19.0, p < 0.001, d = 2.69$) and rules ($t = 28.6p < 0.001, d = 4.04$) than models of constrained and sufficient capacity in all cases (Fig. 2).

Consistent with prior work on catastrophic forgetting, the constrained and sufficient capacity learning systems exhibited very limited ability to retain prior knowledge when having to learn a new set of interfering associations. The excess capacity systems, however, showed an enhanced ability to retain their knowledge of both episodes ($t = 20.3, p < 0.001, d = 2.87$) and rules ($t = 14.1, p < 0.001, d = 1.99$) (Fig. 1 and 2).

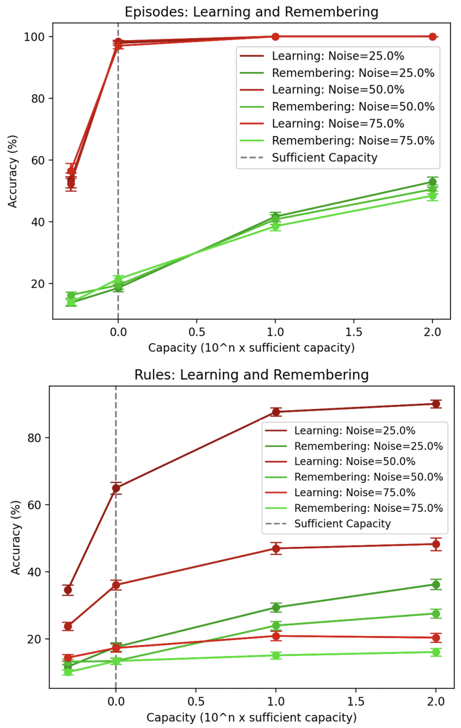

Figure 2: Final averaged mean results after training. Left of the dashed line: constrained capacity; Dashed line: sufficient capacity, Right of the dashed line: excess capacity. Error bars show standard errors.

Conclusion

Our results demonstrate the 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 opposing 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; Sherry & Schacter, 1987; 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

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439.

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