Spotlight

Avoiding Catastrophic Forgetting

Michael E. Hasselmo[1,*]

Humans regularly perform new learning without losing memory for previous information, but neural network models suffer from the phenomenon of catastrophic forgetting in which new learning impairs prior function. A recent article presents an algorithm that spares learning at synapses important for previously learned function, reducing catastrophic forgetting.

As humans go about their daily lives they must constantly learn new information and new skills. For example, when you meet a new acquaintance you must update your memory to subsequently remember their name. When you use new software you must learn the new set of commands for specific functions. Humans regularly transition between new and old tasks, and can easily return to previously learned information when recognizing familiar faces or using older software packages.

The continual learning of different tasks has proved to be a difficult challenge for computational models of neural network models, as addressed in the target article of this Spotlight, Kirkpatrick *et al.* [1]. This problem was recognized at an early stage of neural networks research and was described as the stability–plasticity dilemma [2]. The stability–plasticity dilemma concerns the problem of encountering a new stimulus and preventing the learning of new representations from distorting existing representations. Even in simple neural models of associative memory function, accurate coding of new associations can impair previous associations [3].

The problem of catastrophic interference was noted during the initial wave of research on connectionist modeling of cognition [4]. Catastrophic interference (or catastrophic forgetting) refers to the loss of previous learning during the learning of new information. For example, McCloskey and Cohen [4] modeled a standard learning task involving a list of paired associates, each of which associates a word A with a word B, in one phase of training of a network (using the backpropagation of error algorithm). Subsequently, they trained the network on a list of paired associates coupling each word A with a different word C, and found the previous associations with word B were almost completely lost, in contrast to much better sparing of initial associations in humans [4]. Different types of solutions for this problem have been proposed based on different potential neural mechanisms [3]. Some approaches to the problem evaluate how current input matches with previous input, and, when there is a mismatch of new input with current representations, the current representations are maintained and new representations are formed separately as exceptions [2,3].

One influential solution to this problem has involved the use of two different complementary learning systems [5]. In this account, new information would be rapidly encoded in an episodic memory store, and then this information could be reactivated in an interleaved manner during consolidation of semantic memory, allowing constant updating of the full scope of semantic representations such that new information does not overwrite older information [5]. This required different dynamics for episodic encoding and the interleaved consolidation of memory that could involve dynamical differences between stages of waking and sleep regulated by the neuromodulator acetylcholine [6]. The paper by McClelland *et al*. [5] gave the classic example of learning several examples of birds, each of which flies, resulting in birds being associated with the feature of flying. However, a network that sequentially learns only the fact that penguins are birds that swim can end up shifting the overall category information to indicate that all birds swim. This can be overcome by storing the new information about penguins in episodic memory, and then using a separate consolidation phase to perform interleaved learning of the example of penguins that swim with multiple previous interleaved examples of birds that fly, retaining the dominant category information while incorporating the new information.

With the recent resurgence of focus on connectionist network implementations of cognitive function [7], there has been a resurgence of interest in the problem of catastrophic forgetting. Despite the remarkable successes of deep-learning models in winning visual categorization competitions and performing other cognitive tasks, phenomena such as catastrophic forgetting indicate how distant deep-learning networks still are from effective simulation of artificial general intelligence in humans. Recently, the seminal model of complementary learning systems [5] has been extended and elaborated to address physiological data on replay phenomena within the hippocampal formation and its interactions with semantic representations in the neocortex [8].

In contrast to this systems-level approach related to neurophysiological data, the target article presents an approach using evaluation of the representational significance of individual synaptic weights within a connectionist network [1]. This alternative approach is also supported by biological evidence, such as data indicating that dendritic spines modified by one experience are somehow protected against further modification. In this new approach, individual weights are protected by a mechanism referred to as 'elastic weight consolidation' (EWC). This algorithm protects individual network parameters such as synaptic weights by evaluating their importance for prior learning. Weights are regulated by a quadratic loss function that acts like a spring to pull important weights back toward the previous weight value, thereby selectively reducing the learning rate for the protected synaptic weights while allowing faster learning at less important synaptic weights. The mechanisms of this algorithm need to store the current weight itself, its variance, and its mean. These elements could correspond to the different timecourses of early and late long-term potentiation [9].

The sophisticated dynamics of weight modification in this recent model resemble the use of multiple timescales of synaptic modification as a mechanism for enhancing capacity in neural models of associative memory [10]). This use of multiple timescales corresponds to experimental evidence for multiple timescales of synaptic modification based on different molecular mechanisms, providing the potential for exciting future work linking functional properties of neural models to the detailed molecular mechanisms of synaptic plasticity. In a similar manner, the mechanisms of systems-level consolidation have provided an important theoretical framework for evaluating data from neurophysiological experiments in the context of the formation and long-term stability of memory representations [5,8].

Acknowledgments

The author is supported by the National Institutes of Health (grants R01 MH60013 and R01 MH61492) and by the Office of Naval Research (MURI N00014-16-1-2832).

[1]Center for Systems Neuroscience, Boston University, 2 Cummington Mall, Boston, MA 02215, USA

*Correspondence: hasselmo@bu.edu (M.E. Hasselmo).

http://dx.doi.org/10.1016/j.tics.2017.04.001

References

1. Kirkpatrick, J. *et al.* (2017) Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. U. S. A.* 114, 3521–3526
2. Carpenter, G.A. and Grossberg, S. (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vision Graphics Image Process.* 37, 54–115
3. French, R.M. (1999) Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 128–135
4. McCloskey, M. and Cohen, N.J. (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In *The Psychology of Learning and Motivation* (Bower, G.H., ed.), pp. 109–165, Academic Press
5. McClelland, J.L. *et al.* (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457
6. Hasselmo, M.E. (1999) Neuromodulation: acetylcholine and memory consolidation. *Trends Cogn. Sci.* 3, 351–359
7. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
8. Kumaran, D. *et al.* (2016) What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* 20, 512–534
9. Clopath, C. *et al.* (2008) Tag-trigger-consolidation: a model of early and late long-term-potentiation and depression. *PLoS Comput. Biol.* 4, e1000248
10. Benna, M.K. and Fusi, S. (2016) Computational principles of synaptic memory consolidation. *Nat. Neurosci.* 19, 1697–1706