Active and Selective Forgetting in Artificial Neural Networks

Solim LeGris

COMP596: Brain Inspired AI

McGill University

1. Introduction

The novelist Jorge Luis Borges was prescient in his understanding of the importance of forgetting. In the short story *Funes the Memorious*, the protagonist Funes has an accident and subsequently develops infinite memory. Every moment of his life is now recorded perfectly. Unfortunately, this newfound ability is an impediment, not a superpower. Funes is unable to abstract from experience since every moment of his life and every object he encounters is infinitely unique. Thought requires storing the essential and discarding the unessential. Funes was incapable of thought as a result of his "perfect" memory.

Although additional knowledge may enhance performance on a task, this relation starts decaying as too much information is acquired leading to an opposite effect [Markovitch and Scott, 1993]. Indeed, the rare individuals with abnormally accurate memory showcase this point. The Russian journalist known as Patient S. was extensively studied by Alexander Luria and had documented impairments in tasks mundane to the average person [Luria, 1987]. Traditionally, neuroscience has treated forgetting as secondary but in the last decade or so it has become increasingly evident that forgetting is 1) not a dysfunction of memory and 2) crucial for cognition [Hardt et al., 2013]. Similarly, AI research has historically not paid much attention to forgetting mechanisms other than trying to avoid the well-known phenomenon of catastrophic forgetting (CF). It will be argued here, on the basis of recent advancements in the neurobiology of forgetting, that if AI is to approach human cognitive capacity, it will need to implement *active* and *controlled* forgetting, especially for continual and sequential learning. In Section 1, theories and evidence for active forgetting mechanisms in biological neural networks are discussed while in Section 2, an overview of learning algorithms which broadly integrate ideas from the neuroscience of memory is elaborated. Lastly, a tentative and speculative framework for implementing the ability to learn sequentially without CF in ANNs is discussed in light of the forgetting mechanisms of brains in Section 3.

2. The Neurobiology of Forgetting

The seat of memory at the cellular level is thought to be the synapse. Initially when learning occurs, synapses that are excited are subject to morphological changes that lead to what is known as long-term potentiation (LTP) through Hebbian learning [Hebb, 1949]. Co-activation of neurons is thought to lead to strengthening of synapses in turn forming neural networks supporting memory traces [Josselyn et al., 2015]. Although LTP induction is an experimental procedure, it has presented itself as an ideal candidate mechanism for memory formation because of its persistence [Abraham and Williams, 2003], associative nature and input-specificity [Doherty et al., 2009]. Following encoding, consolidation (i.e. stabilization) occurs both at the synaptic and systems levels through molecular processes such as GluA2-containing AMPA receptor (GluA2/AMPARs) upregulation and hippocampus-dependent mechanisms [Nadel and Hardt, 2011].

Molecular turnover entails that there must be mechanisms other than the brief consolidation events that support memory maintenance [Crick, 1984]. A plausible candidate for LTP persistence was found to be PKMζ, an atypical kinase that is a member of the PKC kinase family [Sacktor, 2011]. Evidence points to a role of PKMζ in preventing endocytosis of GluA2/AMPARs thereby maintaining memories [Migues et al., 2010]. How then is anything forgotten if active mechanisms continuously maintain consolidated memories?

The tendency for memories to be forgotten following Ebbinghaus' forgetting curve was replicated in humans [Murre and Dros, 2015] and disrupted through blocking of GluA2/AMPAR removal in mice [Migues et al., 2016]. Furthermore, [Migues et al., 2016] provided evidence for a forgetting mechanism that was 1) *not* mediated by interference and 2) supported by *active* decay. In other words, forgetting is not merely a passive time-dependent phenomenon but an *active* process that counteracts maintenance mechanisms. Clearly, if maintenance mechanisms are supported by active and dynamic processes, so should forgetting. A more radical hypothesis of memory researchers is that by default the brain forgets [Hardt et al., 2013]. As a result, only what has been “tagged” as important is actively kept.

Why should we have evolved to forget more than we remember? Given the evidence discussed above, it seems that forgetting is a *function* of memory. Arguably, this capacity to parse out irrelevant information and noise is what allows organisms to have flexible and adaptive behaviour [Richards and Frankland, 2017]. For instance, active forgetting may allow brains to generalise over experiences. It was shown that mice trained in a contextual fear paradigm exhibit a freezing response (i.e. fear) in a novel but similar environment when not exposed to the encoding context for 28 days [Lacy and Stark, 2013]. In other words, it seems that over time details fade away only to leave a semanticized memory trace that can be useful in different contexts when relevant. Moreover, it was shown that pharmacologically preventing forgetting in the same contextual fear paradigm did not lead to generalisation when exposed to a new context [Migues et al., 2016]. Previously learnt behaviours often become inadaptive in a constantly changing environment. Consequently, forgetting may support cognitive flexibility as was shown in mice trained in a Morris watermaze task [Awasthi et al., 2019]. Lastly, forgetting must also underlie memory optimization since clearly brains are limited in their ability to store information. Given the mechanisms underlying active decay, brains can optimize and control what information should be retained, for how long and to what extent the stored information is subject to change or destruction [Hardt et al., 2014].

3. Learning Algorithms that Forget

By default, ANNs are implicit forgetting functions. Through the process of learning, they must abstract away details that are irrelevant to the task they are trained to perform [Beierle and Timm, 2019, Markovitch and Scott, 1993, Tishby et al., 2000]. Moreover, broad parallels exist between forgetting in biological neural networks and regularization techniques in ANNs [Richards and Frankland, 2017]. Long short-term memory (LSTM [Hochreiter and Schmidhuber, 1997]) implements attention mechanisms such that a network can make use of present and prior information to generate outputs in a sequential manner. These algorithms also implement "forgetting gates" which allow LSTMs to reset their internal state, preventing breakdown of the network [Gers et al., 1999]. LSTMs implement forgetting using exponential decay, akin to what is described as passive decay in psychological literature. More recently, elastic weight consolidation (ECW) was developed in an attempt to prevent CF by mimicking synaptic consolidation [Kirkpatrick et al., 2017]. The importance of weights (i.e. synapses) is estimated and regularization methods are applied accordingly to prevent too much deviation from previously learnt parameters. In other words, important weights are "frozen" to avoid CF and only allow forgetting of parameter settings that are less important when learning new data. Generative algorithms inspired by the dialogue between the hippocampus (HC) and neocortex have also been developed to implement active forgetting. Dual-memory-based learning systems are inspired by what is known as systems consolidation where the HC stores memory traces for short periods of time and teaches the neocortex memories to be stored long-term, thereby avoiding CF [McClelland et al., 1995]. In order to enable continuous learning, [Sukhov et al., 2020] developed a dual-network model using a feedforward architecture that broadly emulates the neocortex and a variational autoencoder as an artificial HC. This ANN can generate synthetic samples (inspired by hippocampal replay, see [Carr et al., 2011]) from its memory and include those samples with the new dataset to prevent CF. Moreover, active forgetting is induced using synthetic samples of a category and setting their labels to zero, effectively erasing from memory previously learnt knowledge.

The effectiveness of active forgetting was demonstrated in the context of stocks portfolio management [Nakayama and Yoshii, 2000]. In this proposed ANN, "obstacle data" (i.e information causing interference or hindering decision-making) was selectively forgotten by increasing decay rate.

4. Implementing Active and Selective Forgetting

The models explored in Section 3 demonstrate that some preliminary forms of forgetting mechanisms have been implemented in machine learning algorithms and that they have the potential to augment the abilities of learning algorithms. Nonetheless, as is illustrated by the problem of CF, machine learning algorithms forget differently than humans do. Attempts to circumvent CF have been devised using regularization-based approaches [Li and Hoiem, 2017, Kirkpatrick et al., 2017] or dual-memory based learning systems (see [Chen and Liu, 2018] for a review) but the storage-hungry joint training method (i.e. including previously learnt data with new data) seems to remain an upper-bound on sequential learning performance.

Biological learning systems continually adapt to their environments and are not known to be prey to CF. This ability is supported by the leveraging of complex molecular processes, likely including the ones described in Section 2. The standing issue of CF and the inabilility of ANNs to learn sequentially in an efficient manner may be due to the lack of controlled and active forgetting mechanisms similar to the ones found in biological systems. Dual-memory-based learning systems that adopt generative methods are promising since they eliminate the need to store old data and are closer to what biological brains do, especially with respect to replay or reactivation and its involvement in learning/forgetting [Nader and Einarsson, 2010]. A key component that allows brains to learn in a sequential manner is the separation of learning systems into an autoassociative rapid learning system, the HC, and a slow, supervised learning system, the neocortex. Although the former cannot be subject to interference, token of its ability to form non-overlapping representations, the latter is subject to interference because of overlap in representations stored [Hardt et al., 2013]. Interference in the neocortex is kept minimal by allowing the HC to teach it associations, slowly leading to cross-linking of sensory representations. Active decay is thought to occur in the HC whereby memory traces are actively and selectively erased over time while the neocortex learns the representations stored in the HC through reactivation and stores them over longer periods of time. Furthermore, biological brains can modulate the plasticity of certain memories (similarly to what EWC does) depending on factors like relevance and recency. To endow ANNs with sequential learning abilities, it may be necessary to implement 1) a dual-memory-based learning architecture, 2) generative properties to support spontaneous reactivation for continuous knowledge update, 3) use and importance dependent plasticity on acquired representations and 4) active and controlled forgetting dependent on plasticity (3) and stability of the environment. Such learning algorithms may be trained on conventional datasets (i.e images, speech, language, etc) and leverage synthetic samples generated similarly to spontaneous reactivation patterns in biological brains.

5. Conclusion

In conclusion, a brief overview of forgetting mechanisms in the brain as well as in AI were discussed. Given the efficiency and performance of biological neural networks, reverse-engineering their mechanisms may lead to the development of more flexible and capable ANNs. Specifically, it was argued here that active and controlled forgetting mechanisms may be crucial to endow ANNs with capacities similar to those of humans, especially with regards to cognitive flexibility, lifelong learning and generalisation. Neurobiological research on forgetting is still in its infancy but recent advancements have shown that forgetting mechanisms are essential components of memory systems and more broadly, cognitive systems. Further research in ANN should therefore make use of these findings to ameliorate the current state of AI research and progress towards more powerful and efficient machine learning algorithms.

References

[Abraham and Williams, 2003] Abraham, W. C. and Williams, J. M. (2003). Properties and mechanisms of ltp maintenance. *The Neuroscientist*, 9(6):463–474. PMID: 14678579.

[Awasthi et al., 2019] Awasthi, A., Ramachandran, B., Ahmed, S., Benito, E., Shinoda, Y., Nitzan, N., Heukamp, A., Rannio, S., Martens, H., Barth, J., et al. (2019). Synaptotagmin-3 drives ampa receptor endocytosis, depression of synapse strength, and forgetting. *Science*, 363(6422).

[Beierle and Timm, 2019] Beierle, C. and Timm, I. J. (2019). Intentional forgetting: An emerging field in ai and beyond.

[Carr et al., 2011] Carr, M. F., Jadhav, S. P., and Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, 14(2):147.

[Chen and Liu, 2018] Chen, Z. and Liu, B. (2018). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207.

[Crick, 1984] Crick, F. (1984). Neurobiology: Memory and molecular turnover. *Nature*, 312(5990):101–101.

[Doherty et al., 2009] Doherty, A., Fitzjohn, S., and Collingridge, G. (2009). Long-term potentiation (ltp): Nmda receptor role. In Squire, L. R., editor, *Encyclopedia of Neuroscience*, pages 555–560. Academic Press, Oxford.

[Gers et al., 1999] Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

[Hardt et al., 2013] Hardt, O., Nader, K., and Nadel, L. (2013). Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, 17(3):111–120.

[Hardt et al., 2014] Hardt, O., Nader, K., and Wang, Y.-T. (2014). Glua2-dependent ampa receptor endocytosis and the decay of early and late long-term potentiation: possible mechanisms for forgetting of short-and long-term memories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1633):20130141.

[Hebb, 1949] Hebb, D. O. (1949). The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology*, 62:78.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Josselyn et al., 2015] Josselyn, S. A., Köhler, S., and Frankland, P. W. (2015). Finding the engram. *Nature Reviews Neuroscience*, 16(9):521–534.

[Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[Lacy and Stark, 2013] Lacy, J. W. and Stark, C. E. (2013). The neuroscience of memory: implications for the courtroom. *Nature Reviews Neuroscience*, 14(9):649–658.

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

[Luria, 1987] Luria, A. R. (1987). *The Mind of a Mnemonist: A Little Book about a Vast Memory, With a New Foreword by Jerome S. Bruner*. Harvard University Press.

[Markovitch and Scott, 1993] Markovitch, S. and Scott, P. D. (1993). Information filtering: Selection mechanisms in learning systems. *Machine Learning*, 10(2):113–151.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (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. *Psychological review*, 102(3):419.

[Migues et al., 2010] Migues, P. V., Hardt, O., Wu, D. C., Gamache, K., Sacktor, T. C., Wang, Y. T., and Nader, K. (2010). Pkmζ maintains memories by regulating glur2-dependent ampa receptor trafficking. *Nature neuroscience*, 13(5):630.

[Migues et al., 2016] Migues, P. V., Liu, L., Archbold, G. E., Einarsson, E. Ö., Wong, J., Bonasia, K., Ko, S. H., Wang, Y. T., and Hardt, O. (2016). Blocking synaptic removal of glua2-containing ampa receptors prevents the natural forgetting of long-term memories. *Journal of Neuroscience*, 36(12):3481–3494.

[Murre and Dros, 2015] Murre, J. M. and Dros, J. (2015). Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, 10(7):e0120644.

[Nadel and Hardt, 2011] Nadel, L. and Hardt, O. (2011). Update on memory systems and processes. *Neuropsychopharmacology*, 36(1):251–273.

[Nader and Einarsson, 2010] Nader, K. and Einarsson, E. Ö. (2010). Memory reconsolidation: an update. *Annals of the New York Academy of Sciences*, 1191(1):27–41.

[Nakayama and Yoshii, 2000] Nakayama, H. and Yoshii, K. (2000). Active forgetting in machine learning and its application to financial problems. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5, pages 123–128. IEEE.

[Richards and Frankland, 2017] Richards, B. A. and Frankland, P. W. (2017). The persistence and transience of memory. *Neuron*, 94(6):1071–1084.

[Sacktor, 2011] Sacktor, T. C. (2011). How does pkmζ maintain long-term memory? *Nature Reviews Neuroscience*, 12(1):9–15.

[Sukhov et al., 2020] Sukhov, S., Leontev, M., Miheev, A., and Sviatov, K. (2020). Prevention of catastrophic interference and imposing active forgetting with generative methods. *Neurocomputing*, 400:73–85.

[Tishby et al., 2000] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.