



International Journal of Neural Networks and Artificial Intelligence

Journal homepage: <https://trpub.org/journal.php?id=bR2aXTjoAZ>

TRPub

IJNNAI

International Journal of
Neural Networks and
Artificial Intelligence

Vol. 1 Issue 1 2026

Review Article

Open Access

Continual Learning Without Forgetting: Replay, Regularisation, and Architectural Strategies Compared

Yonas Kovacevic^{1*}, Mathieu Khachatryan², Rika Nkosi³

¹Department of Electrical Electronics, Niger Institute for Advanced Electrical Electronics, Indonesia

²Department of Electrical Electronics, Rhine University Centre for Electrical Electronics, Sri Lanka

³Department of Electrical Electronics, Kaveri School of Electrical Electronics, Ethiopia

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijnnai.2026.v1.i1.pp8-14](https://doi.org/10.62715/ijnnai.2026.v1.i1.pp8-14)

Received: 01 October 2025

Accepted: 28 November 2025

Published: 04 February 2026

Abstract

A neural network trained on one task and then on another tends to lose the first. This catastrophic forgetting has been known since the late 1980s, and it stands between current systems and the ability to keep learning after deployment. Three broad families of remedy have emerged. Replay methods keep or regenerate examples of earlier data and mix them into later training. Regularisation methods add a penalty that discourages change in parameters, or in outputs, that mattered for earlier tasks. Architectural methods give each task its own parameters by growing, masking or prompting the network. This narrative review sets out the reasoning behind each family, describes its best-known members, and compares the families on memory cost, computation, dependence on task identity, privacy and behaviour in the harder class-incremental setting. It then turns to evaluation, where the choice of scenario has been shown to change the ranking of methods and where very simple baselines have embarrassed elaborate ones. The comparison suggests that replay is the most dependable family when storing data is permitted, that regularisation alone is seldom enough when task identity is hidden at test time, and that architectural methods built on pretrained backbones have changed what the trade-offs look like.

Keywords: *Continual learning, Catastrophic forgetting, Experience replay, Elastic weight consolidation, Parameter isolation, Class-incremental learning*

© 2026 Yonas Kovacevic et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Most neural networks are trained once, on a fixed dataset shuffled so that every batch resembles every other. The assumption behind stochastic gradient descent is that the examples are drawn independently from a distribution that does not move. Outside the laboratory the distribution does move. A diagnostic model meets a new scanner, a recommender meets new products, a robot is asked to do something it was not built for. Retraining from scratch on everything seen so far is the safe response and often an impractical one, whether for reasons of cost, of time, or because the old data may no longer be held.

Training on the new data alone invites a failure first described by [McCloskey and Cohen \(1989\)](#), who

* Corresponding author: Yonas Kovacevic, Department of Electrical Electronics, Niger Institute for Advanced Electrical Electronics, Indonesia.
E-mail: yonas.kovacevic@niger.edu

© 2026 Yonas Kovacevic et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

found that a network taught a second set of arithmetic facts abruptly lost the first. The effect, reviewed a decade later by French (1999), follows from the distributed way in which networks store what they know: the same weights serve many inputs, so adjusting them for new inputs disturbs the old. The tension between remaining able to learn and retaining what has been learned is usually called the stability-plasticity dilemma (Parisi et al., 2019).

The modern literature on the problem dates from about 2016 and has grown quickly. Its proposals are conventionally sorted into three families: replay, regularisation and architectural or parameter-isolation methods (De Lange et al., 2022). This article reviews the three and compares them, with particular attention to the conditions under which each can be expected to work.

2. Scope and Approach

The article is a narrative review of the published literature on continual learning in deep neural networks, confined largely to supervised classification, where most of the comparative work has been done. It relies on the papers that introduced the principal methods and on surveys and critical studies that compare them. No formal search protocol was followed and no count of sources is claimed; methods were included because they are widely used as reference points. Reinforcement learning, continual pretraining of language models and online learning from a single pass over a stream are mentioned only where they bear on the comparison.

3. Settings and Measures

Comparison is impossible without agreement on what is being attempted, and for some years the field lacked it. van de Ven and Tolias (2019) distinguished three scenarios that are now standard. In task-incremental learning the model is told at test time which task an input belongs to and need only choose among that task's classes. In domain-incremental learning the label set stays the same while the input distribution shifts, and task identity is not given. In class-incremental learning new classes arrive over time and the model must discriminate among all classes seen so far without being told the task. The third is much the hardest, because the network must separate classes it has never seen together.

Performance is usually reported as the average accuracy over all tasks after the last has been learned. Lopez-Paz and Ranzato (2017) added two measures that have been widely adopted: backward transfer, the effect of later learning on earlier tasks, which is negative when there is forgetting, and forward transfer, the effect of earlier learning on tasks not yet trained. A method that freezes everything it has learned shows no forgetting and no capacity to improve old tasks either, so the measures need to be read together.

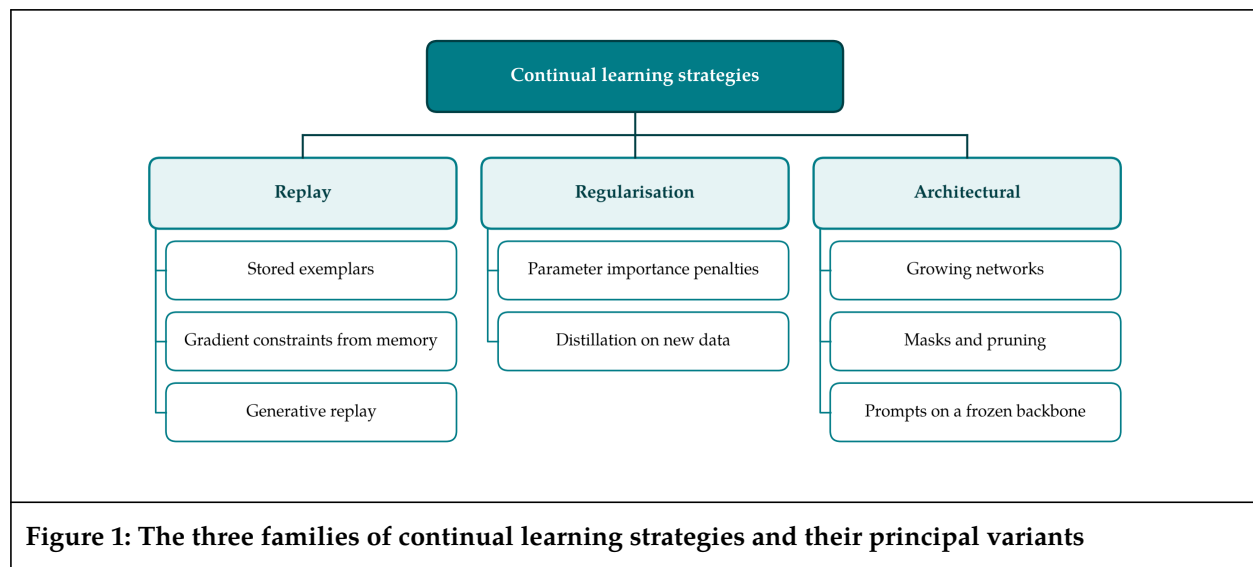
4. Replay Strategies

Replay has the longest pedigree. Robins (1995) showed that interleaving a few old items with new ones protected small networks, and that when old items were unavailable, random inputs labelled by the old network served nearly as well, a procedure he called pseudorehearsal. The idea has a counterpart in the theory of complementary learning systems, according to which the hippocampus records experiences quickly and replays them to a slowly learning neocortex (McClelland et al., 1995). Figure 1 places replay beside the other two families and shows the main variants of each.

In its plainest modern form, a small buffer of past examples is kept and a few are added to every training batch. iCaRL, designed for the class-incremental case, chooses the stored exemplars so that their mean approximates the class mean in feature space, classifies by nearest exemplar mean, and adds a distillation term to hold the representation steady (Rebuffi et al., 2017). Gradient Episodic Memory uses the buffer differently: it treats the losses on stored examples as constraints and projects each update so that none of them increases (Lopez-Paz and Ranzato, 2017). Its successor A-GEM replaces the many constraints with one on the average memory loss and is far cheaper (Chaudhry et al., 2019a).

Two later findings shaped practice. Chaudhry et al. (2019b) reported that ordinary experience replay with a very small memory was a stronger baseline than had been assumed, despite the obvious risk of

overfitting to the few stored items. Dark Experience Replay stores the network's logits alongside each example and replays those, which combines rehearsal with distillation at almost no extra cost (Buzzega et al., 2020). Both results point the same way: simple replay, carefully tuned, is hard to beat.



Where raw data cannot be kept, a generative model can stand in for the buffer. In deep generative replay a generator is trained alongside the classifier and produces synthetic examples of earlier tasks, which are labelled by the previous classifier and mixed with new data (Shin et al., 2017). The approach removes the storage problem and moves the difficulty elsewhere, since the generator must itself learn continually and its errors compound from task to task.

5. Regularisation Strategies

Regularisation methods store no examples. They add to the loss a term that resists change in whatever mattered before. Elastic Weight Consolidation, the best known, takes a Bayesian view: the parameters found for an old task define an approximate posterior, and a quadratic penalty weighted by the diagonal of the Fisher information matrix anchors the weights the old task relied on while leaving the rest free (Kirkpatrick et al., 2017). Synaptic Intelligence reaches a similar penalty by accumulating, during training, each weight's contribution to the fall in loss (Zenke et al., 2017). Memory Aware Synapses estimates importance from the sensitivity of the network's output to each weight, which needs no labels and can be computed on unlabelled data (Aljundi et al., 2018).

A second branch regularises the function and not the weights. Learning without Forgetting records the old network's responses to the new task's inputs before training begins and then asks the updated network to reproduce them, using the distillation loss of Hinton et al. (2015) as the penalty (Li and Hoiem, 2018). No old data are required, only the new data and the old model. The method works best when the new inputs resemble the old; when they do not, the old network's outputs on them are a poor record of what it knew.

The attraction of the family is economy. Memory overhead is at most a few copies of the parameters, there is nothing sensitive to store, and the methods drop into an existing training loop. The weakness is that importance estimates are local approximations made at the end of each task, and their errors accumulate over a long sequence. The penalties also protect each task separately and give the network no signal about how classes from different tasks relate to one another.

6. Architectural Strategies

If interference arises because tasks share weights, one remedy is to stop them sharing. Progressive Neural Networks add a fresh column of layers for every task, freeze the earlier columns, and connect them laterally so that the new column can reuse old features (Rusu et al., 2016). Forgetting is excluded by

construction, and the price is a network that grows with every task. Dynamically Expandable Networks moderate the growth by retraining selectively and adding units only when the existing ones cannot accommodate a task (Yoon et al., 2018).

Other methods keep the network at a fixed size and partition it. PackNet trains a task, prunes the weights that task can spare, and hands the freed weights to the next, recording a binary mask for each (Mallya and Lazechnik, 2018). Hard Attention to the Task learns the masks during training as gates over units (Serrà et al., 2018). In both, the right mask has to be selected at test time, so task identity must be supplied or inferred, and capacity is eventually exhausted.

Large pretrained backbones have given the family a new form. Learning to Prompt keeps a pretrained vision Transformer frozen and learns a small pool of prompt vectors, from which a few are selected for each input by matching against learned keys (Wang et al., 2022). Nothing in the backbone can be overwritten, no rehearsal buffer is needed, and the choice of prompt replaces the explicit task label. The approach depends heavily on the quality of the pretrained representation and on how far the new data lie from what the backbone has seen.

7. The Three Families Compared

Table 1 sets the three families against the criteria that most often decide the choice in practice. The entries describe tendencies of each family, and individual methods depart from them.

The clearest dividing line is the scenario. In task-incremental benchmarks, where the test-time label of the task is known, methods from all three families do well and architectural methods can eliminate forgetting altogether. In the class-incremental scenario van de Ven and Tolias (2019) found that regularisation methods such as Elastic Weight Consolidation and Synaptic Intelligence failed, while methods using some form of replay continued to perform. The broad experimental survey of De Lange et al. (2022), conducted in the task-incremental setting, likewise found that outcomes depend strongly on model capacity, task order and the tuning of the trade-off between stability and plasticity.

Table 1: Replay, regularisation and architectural strategies compared on practical criteria			
Criterion	Replay	Regularisation	Architectural
What is stored	Past examples, logits or a generator	Importance values or an old model copy	Per-task columns, masks or prompts
Growth with tasks	Fixed buffer, or generator cost	Roughly constant	Grows, or capacity runs out
Task identity at test	Not needed	Usually needed for good results	Needed or must be inferred
Class-incremental use	Strongest of the three	Weak on its own	Possible with task inference or prompts
Privacy of old data	A concern unless generative	No raw data kept	No raw data kept
Typical failure	Overfits or drifts with a small buffer	Errors accumulate over long sequences	Cannot improve frozen old tasks
Representative method	Rebuffi et al., 2017	Kirkpatrick et al., 2017	Mallya and Lazechnik, 2018

The families are also less separate than the taxonomy implies. iCaRL combines exemplars with distillation. Dark Experience Replay is replay of a regularisation target. Gradient Episodic Memory uses a buffer to constrain optimisation and never trains on the buffer directly. Many strong systems add a small memory to a penalty or a mask. A practitioner choosing among families is therefore usually choosing a starting point, to which elements of the others can be added as constraints allow. The useful questions are concrete ones: whether old examples may lawfully be kept, how much memory the device can spare, and

whether anything at test time reveals which task an input belongs to.

8. Evaluation and Its Pitfalls

Several critical studies have argued that reported progress depends on how the test is set. [Farquhar and Gal \(2018\)](#) observed that popular benchmarks, including sequences of permuted images, bear little relation to realistic change and that methods which look effective when task identity is given can fail when it is withheld. They proposed that evaluations use more than two tasks, hide task identity at test time, and restrict access to old data.

[Prabhu et al. \(2020\)](#) made the point with a deliberately naive baseline. Their method, GDumb, greedily fills a class-balanced memory as data arrive and, when asked to predict, trains a network from scratch on the memory alone. It makes no attempt to learn continually, and it was nevertheless competitive with many purpose-built methods across the formulations they examined. The result does not show that continual learning is futile. It shows that a method must be compared with what the same memory budget would buy if spent in the simplest way.

9. Gaps and Future Work

The body of comparative evidence is concentrated on image classification with short sequences of clearly separated tasks. Real streams rarely announce their boundaries, and methods that compute importance or freeze masks at the end of a task have no obvious moment at which to do so. Longer sequences, gradual drift and recurring tasks are underexplored. Reporting is uneven as well: memory and computation are often left out of results tables, although they are what a deployed system must budget for.

Pretrained backbones raise a separate question. When most of the representation is learned in advance and frozen, it is unclear how much of the measured resistance to forgetting belongs to the continual method and how much to the pretraining, and benchmarks drawn from data similar to the pretraining set may flatter the approach. A related need is for theory that predicts when a given family will fail; at present the field knows this mostly by trial. Regulatory demands to delete personal data on request also sit awkwardly with replay buffers and deserve technical attention.

10. Conclusion

More than three decades after catastrophic interference was first described, the remedies fall into three recognisable families whose strengths are complementary. Replay is the most reliable where examples may be stored and remains the reference point in class-incremental learning. Regularisation is the lightest to deploy and is adequate where task identity is known and sequences are short. Architectural methods offer the firmest guarantees against forgetting at the price of growth or a need for task inference, and prompts on frozen backbones have made them newly practical. The most useful recent contributions have been critical ones, which insist on declared scenarios, honest baselines and reported costs.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M. and Tuytelaars, T. (2018). [Memory aware synapses: Learning what \(not\) to forget. *Proceedings of the European Conference on Computer Vision \(ECCV\)*.](#)
- Buzzega, P., Boschini, M., Porrello, A., Abati, D. and Calderara, S. (2020). [Dark experience for general continual learning: A strong, simple baseline. *Advances in Neural Information Processing Systems*, 33.](#)
- Chaudhry, A., Ranzato, M., Rohrbach, M. and Elhoseiny, M. (2019a). [Efficient lifelong learning with](#)

A-GEM. *Proceedings of the International Conference on Learning Representations*.

- Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H.S. and Ranzato, M. (2019b). [On tiny episodic memories in continual learning](#). arXiv preprint arXiv:1902.10486.
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T. (2022). [A continual learning survey: Defying forgetting in classification tasks](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366-3385. <https://doi.org/10.1109/TPAMI.2021.3057446>
- Farquhar, S. and Gal, Y. (2018). [Towards robust evaluations of continual learning](#). arXiv preprint arXiv:1805.09733.
- French, R.M. (1999). [Catastrophic forgetting in connectionist networks](#). *Trends in Cognitive Sciences*, 3(4): 128-135. [https://doi.org/10.1016/S1364-6613\(99\)01294-2](https://doi.org/10.1016/S1364-6613(99)01294-2)
- Hinton, G., Vinyals, O. and Dean, J. (2015). [Distilling the knowledge in a neural network](#). arXiv preprint arXiv:1503.02531.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. and Hadsell, R. (2017). [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13): 3521-3526. <https://doi.org/10.1073/pnas.1611835114>
- Li, Z. and Hoiem, D. (2018). [Learning without forgetting](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2935-2947. <https://doi.org/10.1109/TPAMI.2017.2773081>
- Lopez-Paz, D. and Ranzato, M. (2017). [Gradient episodic memory for continual learning](#). *Advances in Neural Information Processing Systems*, 30.
- Mallya, A. and Lazebnik, S. (2018). [PackNet: Adding multiple tasks to a single network by iterative pruning](#). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7765-7773.
- 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): 419-457. <https://doi.org/10.1037/0033-295X.102.3.419>
- McCloskey, M. and Cohen, N.J. (1989). [Catastrophic interference in connectionist networks: The sequential learning problem](#). *Psychology of Learning and Motivation*, 24: 109-165. [https://doi.org/10.1016/S0079-7421\(08\)60536-8](https://doi.org/10.1016/S0079-7421(08)60536-8)
- Parisi, G.I., Kemker, R., Part, J.L., Kanan, C. and Wermter, S. (2019). [Continual lifelong learning with neural networks: A review](#). *Neural Networks*, 113: 54-71. <https://doi.org/10.1016/j.neunet.2019.01.012>
- Prabhu, A., Torr, P.H.S. and Dokania, P.K. (2020). [GDumb: A simple approach that questions our progress in continual learning](#). *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Rebuffi, S.-A., Kolesnikov, A., Sperl, G. and Lampert, C.H. (2017). [iCaRL: Incremental classifier and representation learning](#). *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001-2010.
- Robins, A. (1995). [Catastrophic forgetting, rehearsal and pseudorehearsal](#). *Connection Science*, 7(2): 123-146. <https://doi.org/10.1080/09540099550039318>
- Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. and Hadsell, R. (2016). [Progressive neural networks](#). arXiv preprint arXiv:1606.04671.
- Serrà, J., Surís, D., Miron, M. and Karatzoglou, A. (2018). [Overcoming catastrophic forgetting with hard attention to the task](#). *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80.
- Shin, H., Lee, J.K., Kim, J. and Kim, J. (2017). [Continual learning with deep generative replay](#). *Advances in Neural Information Processing Systems*, 30.
- van de Ven, G.M. and Tolias, A.S. (2019). [Three scenarios for continual learning](#). arXiv preprint arXiv:1904.07734.

- Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J. and Pfister, T. (2022). [Learning to prompt for continual learning](#). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yoon, J., Yang, E., Lee, J. and Hwang, S.J. (2018). [Lifelong learning with dynamically expandable networks](#). *Proceedings of the International Conference on Learning Representations*.
- Zenke, F., Poole, B. and Ganguli, S. (2017). [Continual learning through synaptic intelligence](#). *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70: 3987-3995.