Reactivation in Artificial Neural Networks

Gido M. van de Ven[1,2] & Andreas S. Tolias[1,2]

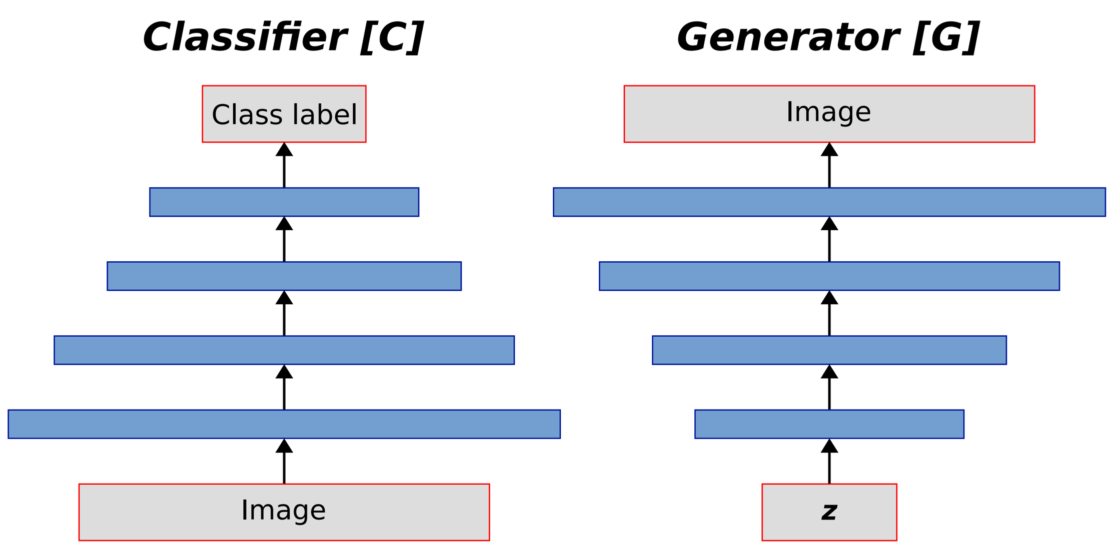
[1] Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, US
[2] Neuroscience Department, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, US

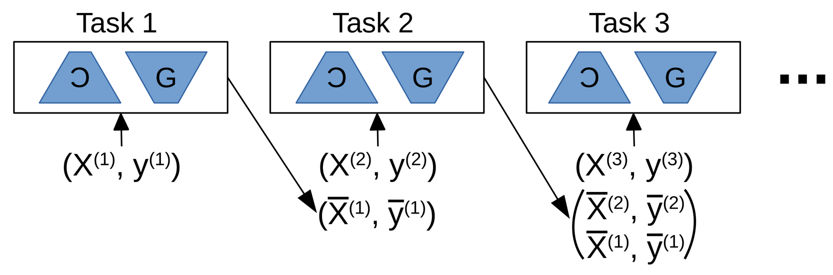
Summary

Current state-of-the-art artificial neural networks can be trained to achieve great performance on a wide variety of individual tasks. However, when trained on a new task, standard neural networks lose most information related to previously learned tasks, a phenomenon referred to as "catastrophic forgetting". The human brain, in contrast, can continually learn new tasks without such dramatic forgetting of previously acquired information. It is thought that the offlne reactivation of memory-representing cell assembly patterns in the hippocampus is important for this capability [1,2]. Here, we explore the possibillity of adding reactivation to artificial neural networks in order to reduce catastrophic forgetting. We show that generative replay outperforms competing methods on all setups of a task protocol involving classification of MNIST-digits, and we propose two brain-inspired improvements to make this strategy scalable to task protocols with more complicated inputs.

Deep Generative Replay (DGR)

As proposed by [3], along with the classifier, a separate generative model is sequentially trained on all tasks. When training on a new task, images generated by the (previous) generative model are labelled by the (previous) classifier and interleaved with the current task's training data. Both the classifier and the generative model are trained with replay.





Methods to compare against

Sequential: The base model is sequentially trained on all tasks.

Learning without Forgetting (LwF): Following [5], the inputs from the current task are labelled by the previous classifier and replayed (*i.e.*, compared with DGR, the generated input samples are replaced by the current task's inputs).

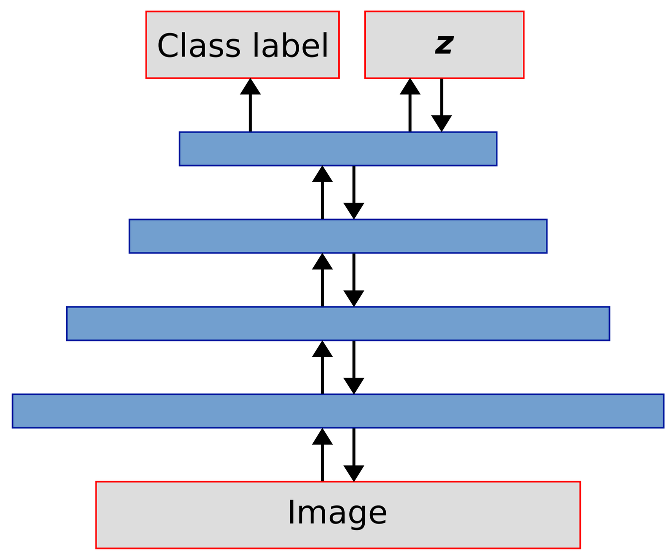
Synaptic Intelligence (SI): Following [6], it is aimed to slow down learning for parameters that are important for previously learned task. To achieve this, a regularization term is added to the loss that penalizes changes to parameters depending on their estimated importance.

Joint: The base model is always trained using the data of all tasks so far.

Brain-inspired Generative Replay

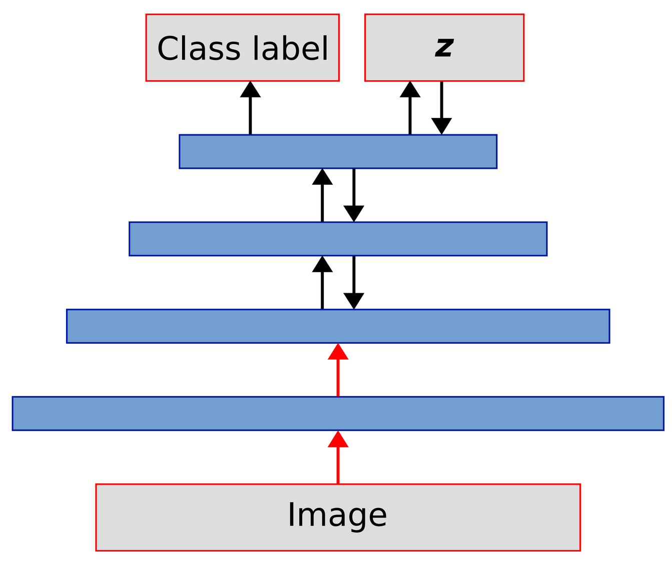
[1] Replay-through-Feedback (RtF)

We integrate the generative model into the classifier by equipping it with generative feedback connections. The resulting model is a symmetric variational auto-encoder [4] with added softmax classification layer.

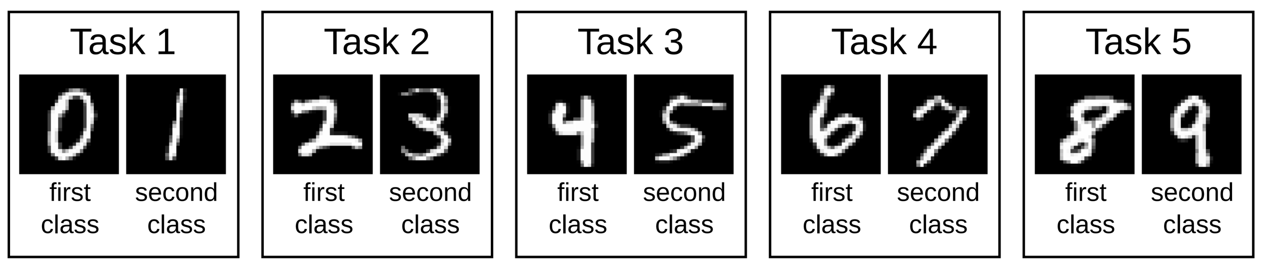


[2] Replay "hidden" representations

Instead of replaying at the pixel level, we replay images at the hidden level. Intuition is that it should be easier to generate such hidden representations. The forward connections that are not replayed are pre-trained and frozen.



Classifying MNIST-digits

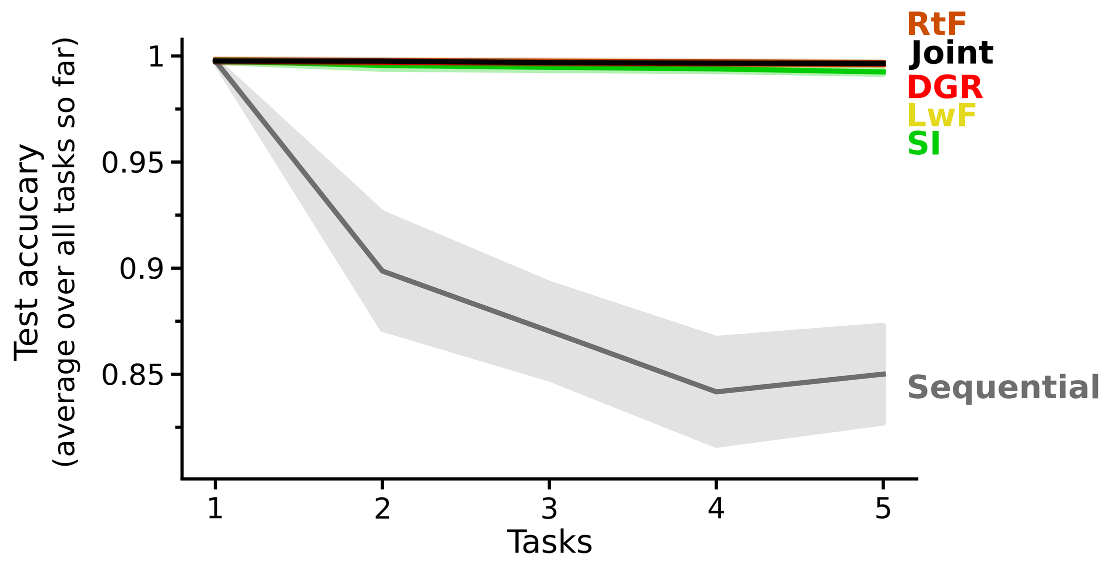


→ Generative replay outperforms competing methods in all setups

→ RtF and DGR perform similarly, with the computational costs of RtF being substantially smaller

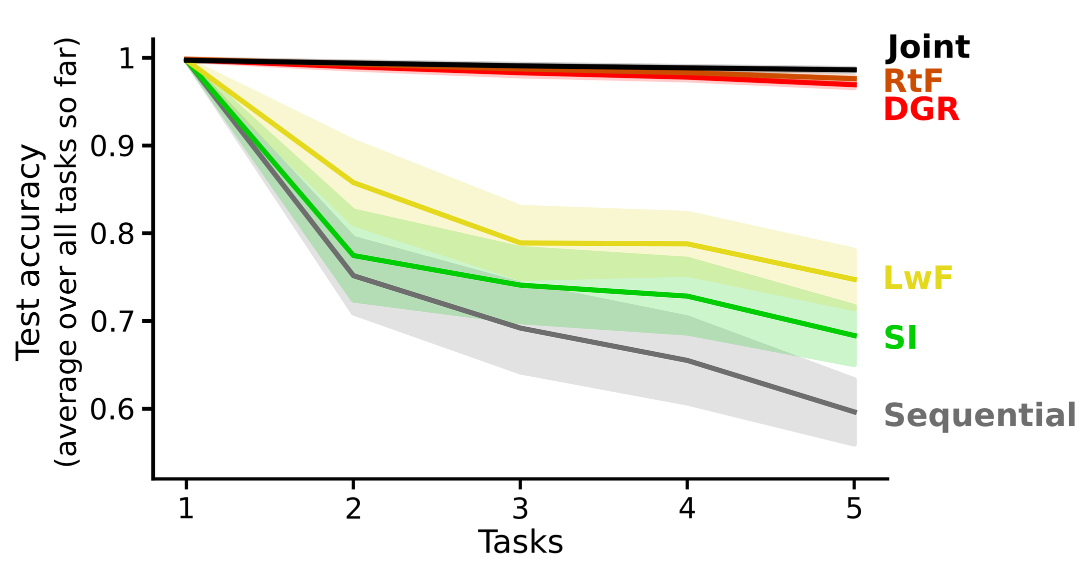
***Setup 1:* Incremental Task Learning**

Task identity is provided. For this task protcol, this means that the choice is always between two known digits (*e.g.*, is it '0' or a '1'?).



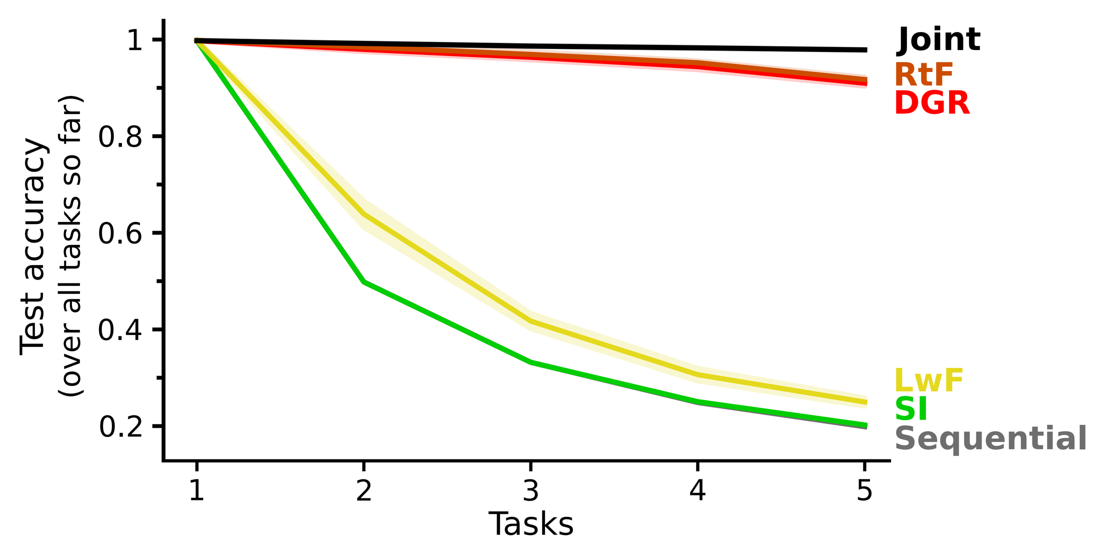
***Setup 2:* Incremental Domain Learning**

Task identity is not provided, but also does not need to be inferred. This means the choice is whether a digit is a "first class" or a "second class" (*e.g.*, is it in ['0', '2', '4', '6', '8'] or in ['1', '3', '5', '7', '9']?).

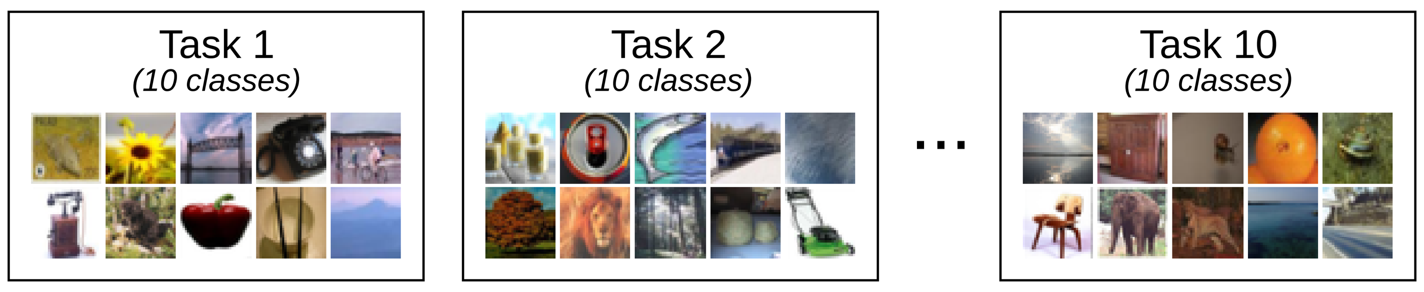


***Setup 3:* Incremental Class Learning**

Task identity is not provided and needs to be inferred as well. This means the choice is between all digits (*i.e.*, choice from '0' to '9').

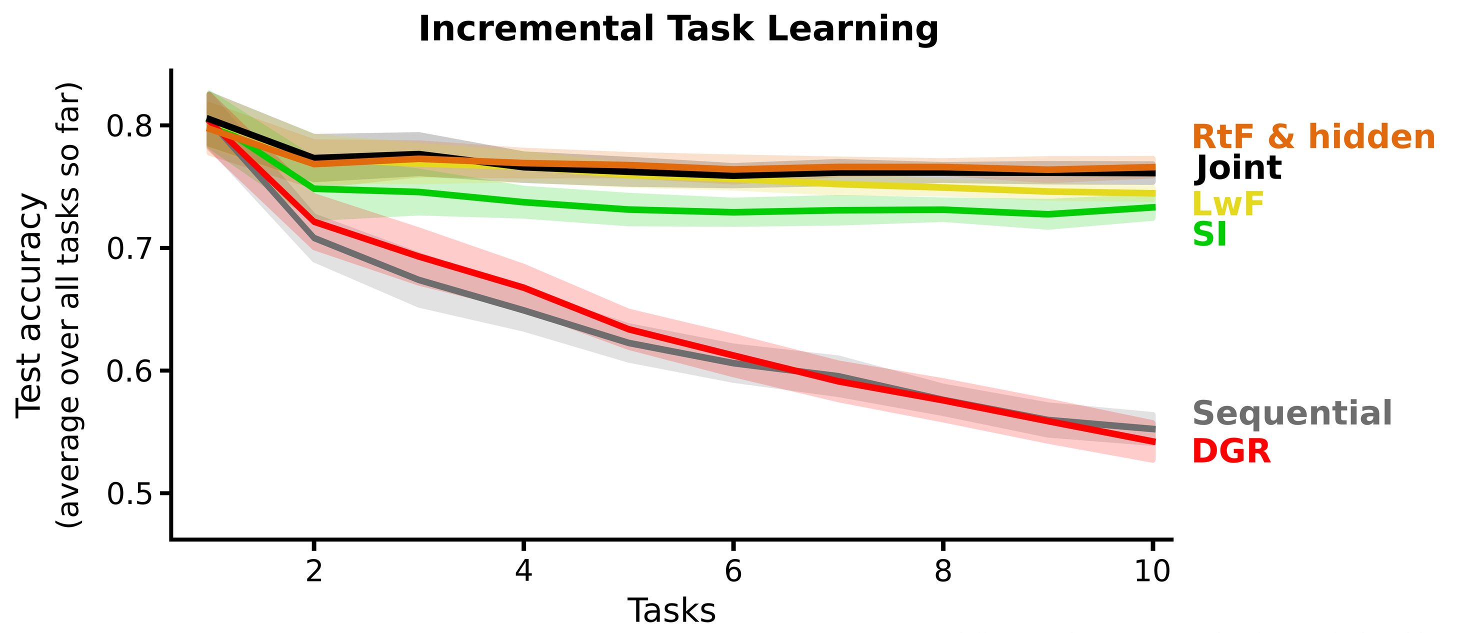


Classifying natural images



→ Straight-forward implementation of generative replay does not scale well to task protocols with more complicated inputs

→ Replaying hidden representations enables generative replay to also be successfull on task protocols with natural images



(all networks use pre-trained convolutional layers)

References

[1] Wilson & McNaughton (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* **265**: 676-679.
[2] van de Ven *et al.* (2016) Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron* **92**: 968-974.
[3] Shin *et al.* (2017) Continual learning with deep generative replay. *NIPS*: 2994-3003.
[4] Kingma & Welling (2014) Auto-encoding variational bayes. *arXiv*: 1412.6980.
[5] Li & Hoiem (2017) Learning without forgetting. *IEEE Trans Pattern Anal Mach Intell*.
[6] Zenke *et al.* (2017) Improved multitask learning through synaptic intelligence. *arXiv*: 1703.04200.

Acknowledgements

This research project is supported by an IBRO-ISN Research Fellowship, by the Lifelong Learning Machines (L2M) program of the Defence Advanced Research Projects Agency (DARPA) via contract number HR0011-18-2-0025 and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.