Three Continual Learning Scenarios

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

Introduction

When trained on a new task, deep neural networks tend to forget most information related to previously learned tasks, a phenomenon referred to as ''catastrophic forgetting''. In recent years, numerous methods for alleviating catastrophic forgetting have been proposed. However, due to the wide variety of experimental protocols used to evaluate them, many of these methods claim ''state-of-the-art'' performance [1,2,3,4,5,6]. To obscure things further, some methods shown to perform well in some experimental settings are reported to dramatically fail in others: compare the performance of elastic weight consolidation in [1] and [7] with that in [8] and [9].

TL;DR: To enable fair and structured comparisons of methods for continual learning, we identify three distinct continual learning scenarios of increasing difficulty.

Continual learning scenarios

[1] Incremental task learning:
This is the easiest continual learning scenario, as models are always informed about which task needs to be performed. Since task identity is always provided, it is possible to train models with task-specific components. A typical neural network architecture used in this scenario has a ''multihead'' output-layer, meaning that each task has its own output units but the rest of the network is (potentially) shared between tasks.

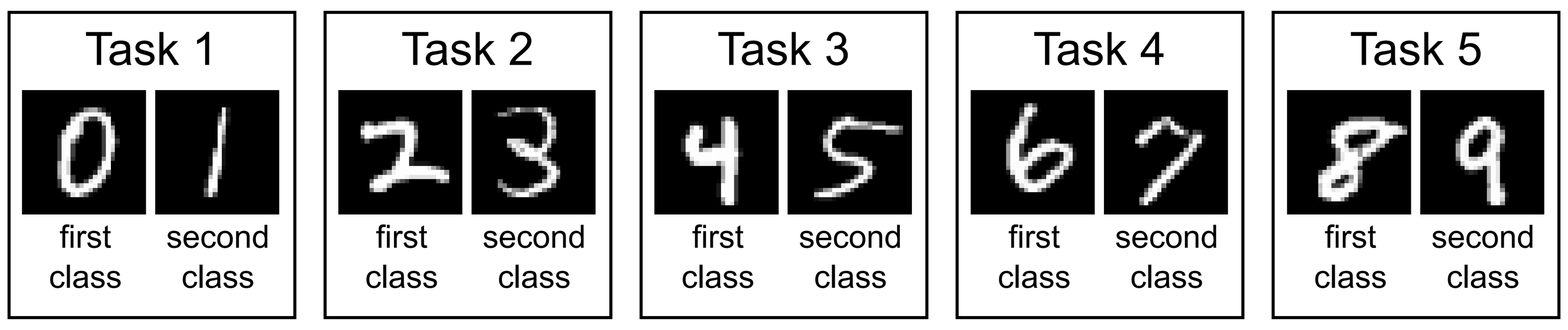
[2] Incremental domain learning:
In this scenario, task identity is not available at test time. Models however only need to solve the task at hand; they are not required to infer which task it is. Typical examples of this scenario are protocols whereby the structure of the tasks is always the same, but the input-distribution is changing.

[3] Incremental class learning:
This is the most difficult scenario, as models need to be able both to solve each task seen so far and to infer which task they are presented with. This scenario includes protocols in which new classes need to be learned incrementally

Compared Methods

- **XdG:** For each task, a random subset of *X*% of the units in each hidden layer are fully gated (i.e., their activations set to zero), with *X* a hyperparameter set by grid search [4].
- **EWC / Online EWC / SI:** A reglarization term is added to the loss that penalizes changes to parameters as a function of how important they are estimated to be for previously learned tasks [1,10,7].
- **LwF:** Images of the current task are replayed with soft targets provided by a copy of the model stored after finishing training on the previous task [11].
- **DGR:** A separate generative model is trained to generate images to be replayed, which are paired with hard targets provided by a copy of the main model stored after training on the previous task [12].
- **DGR+distill:** A separate generative model is trained to generate the images to be replayed, but these are paired with soft targets (as in LwF) instead of hard targets (as in DGR).

Split MNIST

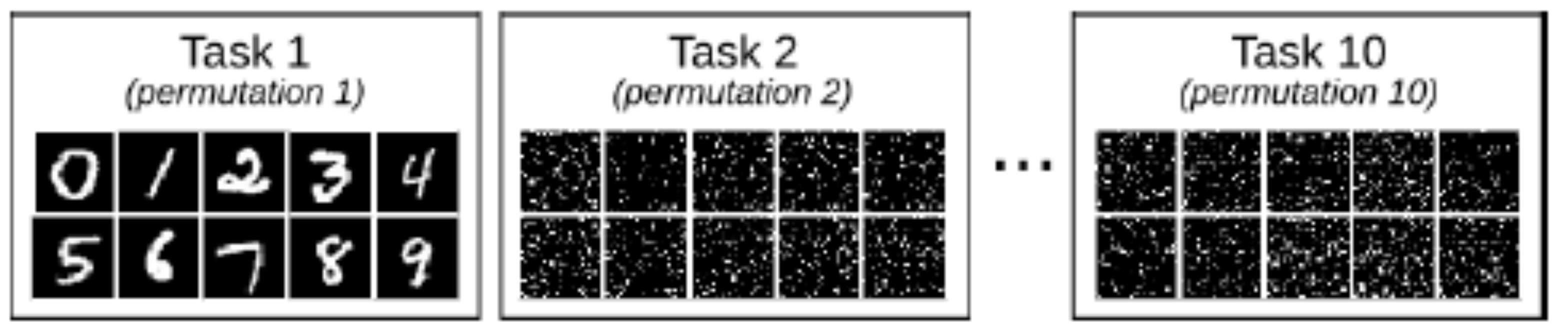


Incremental task learning	With task given, is it the first or second class? (*e.g.*, '0' or '1')
Incremental domain learning	With task unknown, is it a first or second class? (*e.g.*, in ['0','2','4','6','8'] or in ['1','3','5','7','9'])
Incremental class learning	With task unknown, which digit is it? (choice from '0' to '9')

All methods used a multi-layer perceptron with 2 hidden layers of 400 nodes each followed by a softmax output layer, 2000 iterations per task and ADAM-optimizer with learning rate 0.001. Reported is the average task accuracy over all tasks:

Method	Incremental task learning	Incremental domain learning	Incremental class learning
None – *lower bound*	85.15 (± 1.00)	57.33 (± 1.66)	19.90 (± 0.02)
XdG	98.74 (± 0.31)	-	-
EWC	85.48 (± 1.20)	57.80 (± 1.61)	19.90 (± 0.02)
Online EWC	85.22 (± 1.06)	57.60 (± 1.66)	19.90 (± 0.02)
SI	99.14 (± 0.11)	63.77 (± 1.18)	20.04 (± 0.08)
LwF	99.60 (± 0.03)	71.02 (± 1.26)	24.17 (± 0.51)
DGR	99.47 (± 0.03)	95.74 (± 0.23)	91.24 (± 0.33)
DGR+distill	99.59 (± 0.03)	96.94 (± 0.14)	91.84 (± 0.27)
Offline – *upper bound*	99.64 (± 0.03)	98.41 (± 0.06)	97.93 (± 0.04)

Permuted MNIST



Incremental task learning	Given permutation *X* was applied, which digit is it?
Incremental domain learning	With permutation unknown, which digit is it?
Incremental class learning	Which digit is it *and* which permutation was applied?

All methods used a multi-layer perceptron with 2 hidden layers of 1000 nodes each followed by a softmax output layer, 5000 iterations per task and ADAM-optimizer with learning rate 0.0001. Reported is the average task accuracy over all tasks:

Method	Incremental task learning	Incremental domain learning	Incremental class learning
None – *lower bound*	81.79 (± 0.48)	78.51 (± 0.24)	17.26 (± 0.19)
XdG	91.40 (± 0.23)	-	-
EWC	94.79 (± 0.03)	94.43 (± 0.10)	27.65 (± 0.52)
Online EWC	96.09 (± 0.08)	94.25 (± 0.15)	34.41 (± 0.66)
SI	94.75 (± 0.14)	95.33 (± 0.11)	29.31 (± 0.62)
LwF	69.84 (± 0.46)	72.64 (± 0.52)	22.64 (± 0.23)
DGR	92.52 (± 0.08)	95.09 (± 0.04)	92.19 (± 0.09)
DGR+distill	97.51 (± 0.01)	97.35 (± 0.02)	96.38 (± 0.03)
Offline – *upper bound*	97.68 (± 0.01)	97.59 (± 0.01)	97.59 (± 0.02)

Discussion

An important finding is that even for relatively simple task protocols involving the classification of MNIST-digits, regularization-based methods such as EWC, Online EWC and SI completely fail in the incremental class learning scenario (i.e., when task identity needs to be inferred as well). Only the replay-based methods DGR and DGR+distill are capable of performing well in this scenario.

References

[1] Kirkpatrick et al. (2017) Overcoming catastrophic forgetting in neural networks. *PNAS* **114**: 3521-3526.
[2] Rebuffi *et al.* (2016) ICARL: incremental classifier and representation learning. *Proc CVPR*.
[3] Nguyen et al. (2018) Variational continual learning. *ICLR*.
[4] Masse et al. (2018) Alleviating catastrophic forgetting using context-dependent gating. *arXiv*: 1802.01569.
[5] Kemker & Kanan (2018) Fearnet: Brain-inspired model for incremental learning. *ICLR*.
[6] Wu et al. (2018) Incremental classifier learning with generative adverserial networks. *arXiv*: 1802.00853.
[7] Zenke et al. (2017) Improved multitask learning through synaptic intelligence. arXiv: 1703.04200.
[8] Kemker et al. (2017) Measuring catastrophic forgetting in neural networks. *arXiv*: 1708.02072.
[9] Kamra et al. (2017) Deep generative dual memory network for continual learning. *arXiv*: 1710.10368.
[10] Schwarz et al. (2018) Progress & compress: A scalable framework for continual learning. *arXiv*: 1805.06370.
[11] Li & Hoiem (2017) Learning without forgetting. *IEEE Trans Pattern Anal Mach Intell*.
[12] Shin et al. (2017) Continual learning with deep generative replay. *NIPS*: 2994-3003.

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.