Baylor College of Medicine

UNIVERSITY OF CAMBRIDGE

KU LEUVEN

RICE

Three types of incremental learning

Gido M. van de Ven, Tinne Tuytelaars & Andreas S. Tolias

_

First Conference on Lifelong Learning Agents (CoLLAs), Montreal

August 2022

Overview

- Identify three continual learning scenarios
  - Intuitively describe them
  - Define them formally in a restricted, ‘academic’ setting
  - Generalize them to more flexible, ‘task-free’ settings
- Review strategies for continual learning
- Empirically compare these strategies on each scenario

What is continual learning?

- In *classical machine learning*, an algorithm has access to all training data at the same time

- With *continual learning*, two key differences are:
  - the training data arrives incrementally
  - the distribution from which the training data is sampled changes over time

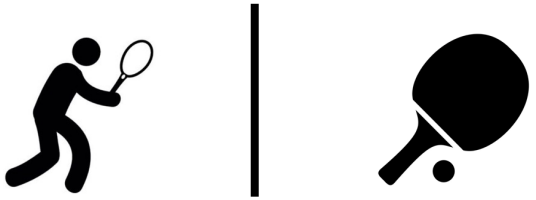
- Main point of this paper:

> A useful way to categorize continual learning problems is based on how the **aspect of the data that changes over time** relates to the **mapping to be learned**

Three continual learning scenarios: intuitively

- Task-incremental learning *(Task-IL)*
  - Incrementally learn a set of clearly distinguishable tasks

Main challenge: achieve positive transfer between tasks

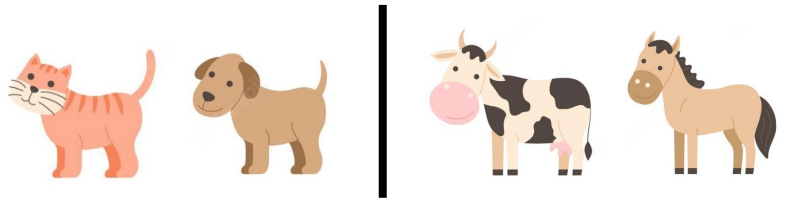


- Domain-incremental learning *(Domain-IL)*
  - Learn the same type of problem in different contexts

Main challenge: alleviate catastrophic forgetting

- Class-incremental learning *(Class-IL)*
  - Incrementally learn a growing number of classes

Main challenge: learn to discriminate between objects not observed together



Images designed by Freepik

See also the preprint: van de Ven & Tolias (2019) Three scenarios for continual learning. *arXiv preprint,* https://arxiv.org/abs/1904.07734

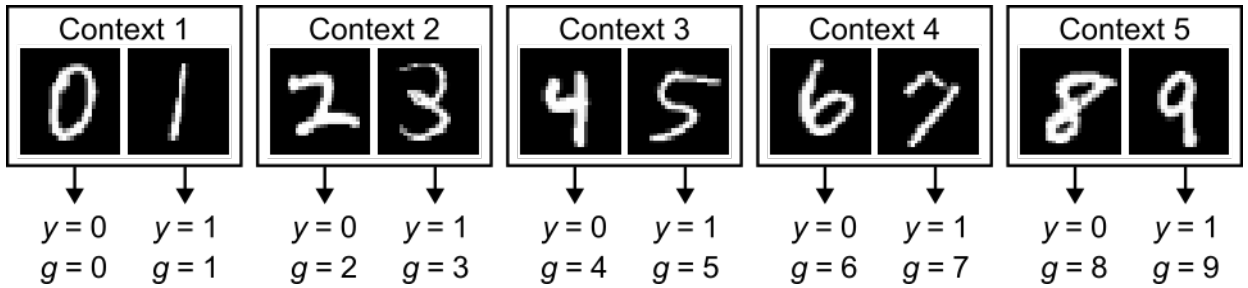
Formalization in “academic continual learning setting”

- Classification problem split in episodes learned sequentially, no overlap between episodes
  - We call these episodes “contexts” (rather than “tasks”)

- Express each sample as consisting of:
  - Input $x \in \mathcal{X}$
  - Within-context label $y \in \mathcal{Y}$
  - Context label $c \in \mathcal{C}$

  } Global label $g \in \mathcal{G}$ (with $\mathcal{G} = \mathcal{C} \times \mathcal{Y}$)

Split MNIST example



- The three scenarios can then be defined based on how the context space $\mathcal{C}$ relates to the mapping to learn:

	Mapping to learn	***Description of choice with Split MNIST***
Task-incremental learning	$f\colon \mathcal{X} \times \mathcal{C} \to \mathcal{Y}$	Choice between two digits of same context (*e.g.* 0 or 1?)
Domain-incremental learning	$f\colon \mathcal{X} \to \mathcal{Y}$	Is the digit odd or even?
Class-incremental learning	$f\colon \mathcal{X} \to \mathcal{C} \times \mathcal{Y}$ or $f\colon \mathcal{X} \to \mathcal{G}$	Choice between all ten digits

Generalization to more flexible settings: theory

- Introduce distinction:
  - **Context set:** collection of underlying distributions, denoted by $\{\mathcal{D}_c\}_{c\in\mathcal{C}}$
  - **Data stream:** sequence of experiences $e_1, e_2, \ldots$ presented to algorithm

- Every observation in the data stream can be sampled from any combination of underlying datasets from the context set:

$$e_t[i] \sim \sum_{c\in\mathcal{C}} p_c^{t,i} \mathcal{D}_c$$

  whereby $e_t[i]$ is observation i of experience t and $p_c^{t,i}$ is the probability that this observation is sampled from $\mathcal{D}_c$.

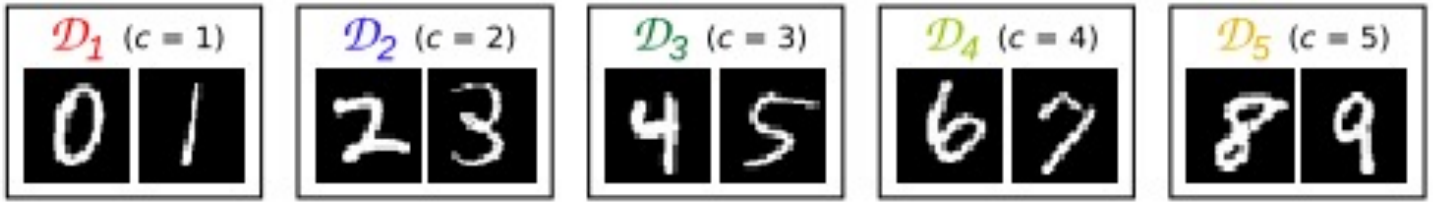
- From a probabilistic perspective, this means two observations at different points in time can only differ w.r.t. the context(s) they are sampled from

  → **the context space $\mathcal{C}$ describes the non-stationary aspect of the data**

- Generalized versions of the three scenarios can be defined as before, based on how the context space $\mathcal{C}$ relates to the mapping to learn.

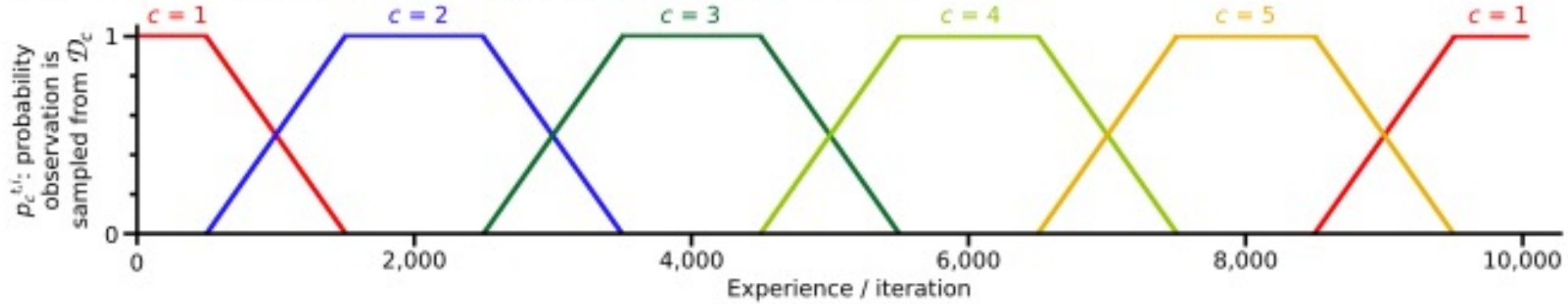
Generalization to more flexible settings: example

[1] Context set — *Specifies what aspect of the data changes over time (i.e., the non-stationary aspect)*



$\mathcal{X}$ = image pixel space
$\mathcal{C}$ = {1,2,3,4,5}
$\mathcal{Y}$ = {0,1}
$\mathcal{G}$ = {0,1,2,3,4,5,6,7,8,9}

[2] Data stream — *Specifies how the non-stationary aspect changes over time*

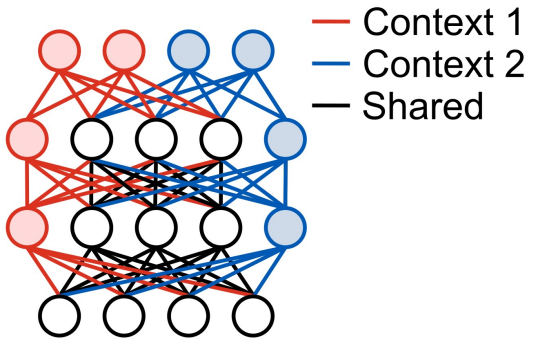


[3] Scenario — *Specifies how the non-stationary aspect relates to the mapping that must be learned*

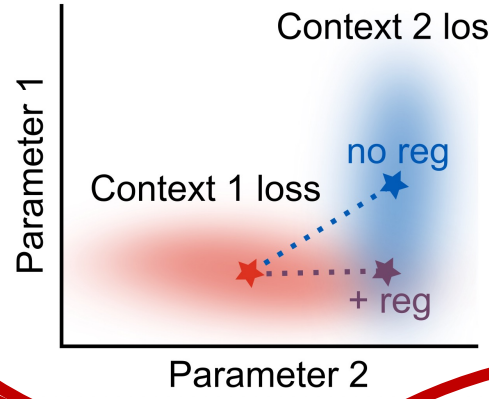
	Type of choice	***Mapping to be learned***
Generalized task-incremental learning	Choice between two digits of same context	$f: \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{Y}$
Generalized domain-incremental learning	Is the digit odd or even?	$f: \mathcal{X} \rightarrow \mathcal{Y}$
Generalized class-incremental learning	Choice between all ten digits	$f: \mathcal{X} \rightarrow \mathcal{C} \times \mathcal{Y}$ or $f: \mathcal{X} \rightarrow \mathcal{G}$

Strategies for continual learning

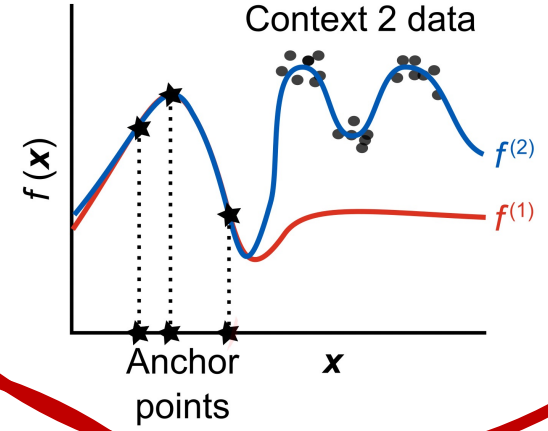
Context-specific components



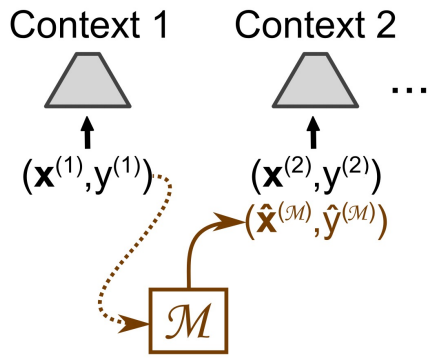
Parameter regularization



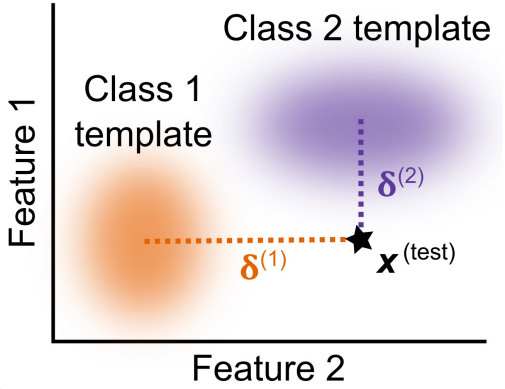
Functional regularization



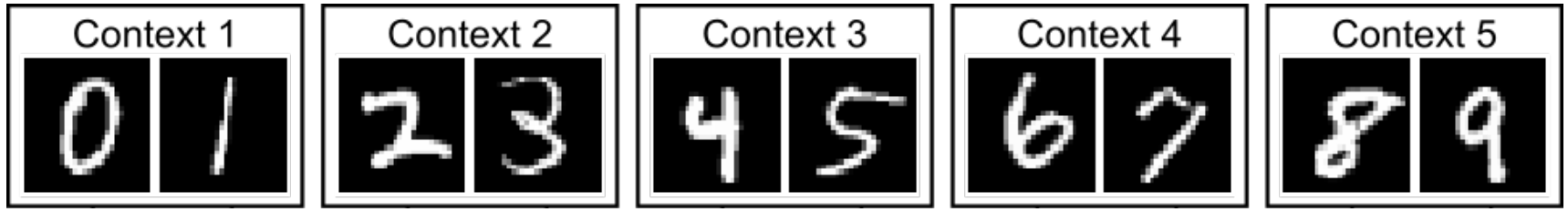
Replay



Template-based classification



Empirical comparison: Split MNIST



Task-incremental learning	Choice between two digits of same context (*e.g.*, 0 or 1?)
Domain-incremental learning	Is the digit odd or even?
Class-incremental learning	Choice between all ten digits

The same sequence of contexts can be “performed” in three different ways:

→ use for a direct comparison between the three scenarios

Empirical comparison: Split MNIST

Strategy	Method	Budget	GM	Task-IL	Domain-IL	Class-IL
Baselines	*None – lower target*			*84.32 (± 0.99)*	*60.13 (± 1.66)*	*19.89 (± 0.02)*
	Joint – upper target			*99.67 (± 0.03)*	*98.59 (± 0.05)*	*98.17 (± 0.04)*
Context-specific components	Separate Networks	-	-	99.57 (± 0.03)	-	-
	XdG	-	-	99.10 (± 0.10)	-	-
Parameter regularization	EWC	-	-	99.06 (± 0.15)	63.03 (± 1.58)	20.64 (± 0.52)
	SI	-	-	99.20 (± 0.11)	66.94 (± 1.13)	21.20 (± 0.57)
Functional regularization	LwF	-	-	99.60 (± 0.03)	71.18 (± 1.42)	21.89 (± 0.32)
	FROMP	100	-	99.12 (± 0.13)	84.86 (± 1.02)	77.38 (± 0.64)
Replay	DGR	-	yes	99.50 (± 0.03)	95.57 (± 0.30)	90.35 (± 0.24)
	BI-R	-	yes	99.61 (± 0.03)	97.26 (± 0.15)	94.41 (± 0.15)
	ER	100	-	98.98 (± 0.07)	93.75 (± 0.24)	88.79 (± 0.20)
	A-GEM	100	-	98.54 (± 0.10)	87.67 (± 1.33)	65.10 (± 3.64)
Template-based classification	Generative Classifier	-	yes	-	-	93.82 (± 0.06)
	iCaRL	100	-	-	-	92.49 (± 0.12)

Shown is final test accuracy (as %, averaged over all contexts). Academic continual learning setting was used. 'Budget' indicates number of samples per class stored in memory, 'GM' indicates generative model was learned using extra parameters. Experiments were run 20 times, reported is mean (± SEM). ***More comparisons in the paper: Split CIFAR-100 and a 'task-free' version of Split MNIST.***

PyTorch code for all experiments: https://github.com/GMvandeVen/continual-learning

For method abbreviations and references, see extra slide.

Summary

- Continual learning is not a unitary problem: we describe three fundamentally different scenarios, each with their own challenges

- These scenarios differ substantially in terms of difficulty and in terms of the effectiveness of different computational strategies

Update relative to preprint version (van de Ven & Tolias, 2019; *arXiv*):

- We formally define these scenarios in a restricted, ‘academic’ continual learning setting; and we generalize them to more flexible, ‘task-free’ settings

Funding acknowledgements

This research project has been supported by an IBRO-ISN Research Fellowship, by the ERC-funded project KeepOnLearning (reference number 101021347), 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. 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 DARPA, IARPA, DoI/IBC, or the U.S. Government.

Abbreviations and references of compared methods

- Context-specific components
  - ***Context-dependent Gating*** **(XdG)**
    Masse NY, Grant GD, Freedman DJ (2018) Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *PNAS* **115**: E10467-E10475.
- Parameter regularization
  - ***Elastic Weight Consolidation*** **(EWC)**
    Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, …, Hadsell R (2017) Overcoming catastrophic forgetting in neural networks. *PNAS* **114**: 3521-3526.
  - ***Synaptic Intelligence*** **(SI)**
    Zenke F, Poole B, Ganguli S (2017) Continual learning through synaptic intelligence. *ICML*: 3987-3995.
- Functional regularization
  - ***Learning without Forgetting*** **(LwF)**
    Li Z, Hoiem D (2017) Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40***:* 2935-2947.
  - ***Functional Regularization Of Memorable Past*** **(FROMP)**
    Pan P, Swaroop S, Immer A, Eschenhagen R, Turner R, Khan ME (2020) Continual deep learning by functional regularisation of memorable past. *NeurIPS*: 4453-4464.
- Replay
  - ***Deep Generative Replay*** **(DGR)**
    Shin H, Lee JK, Kim J, Kim J (2017) Continual learning with deep generative replay. *NeurIPS*: 2994-3003.
  - ***Brain-Inspired Replay*** **(BI-R)**
    van de Ven GM, Siegelmann HT, Tolias AS (2020) Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications* **11**: 4069.
  - ***Experience Replay*** **(ER)**
    Rolnick D, Ahuja A, Schwarz J, Lillicrap T, Wayne G (2019) Experience replay for continual learning. *NeurIPS*: 32
    Chaudhry A, Rohrbach M, Elhoseiny M, Ajanthan T, Dokania PK, Torr PH, Ranzato MA (2019) On tiny episodic memories in continual learning. *arXiv preprint*: 1902.10486.
  - ***Averaged Gradient Episodic Memory*** **(A-GEM)**
    Chaudhry A, Ranzato MA, Rohrbach M, Elhoseiny M (2019) Efficient Lifelong Learning with A-GEM. *ICLR*.
- Template-based classification
  - **Generative Classifier**
    van de Ven GM, Zhe L, Tolias AS (2021) Class-incremental learning with generative classifiers. *CVPR-W proceedings*: 3611-3620.
  - ***Incremental Classifier and Representation Learning*** **(iCaRL)**
    Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH (2017) icarl: Incremental classifier and representation learning. *CVPR proceedings*: 2001-2010.