Séminaire IA Univ Eiffel

Distillation in Online Continual Learning

Nicolas MICHEL
4th year PhD student
LIGM

nicolas1203.github.io

September 19, 2024

Table of content

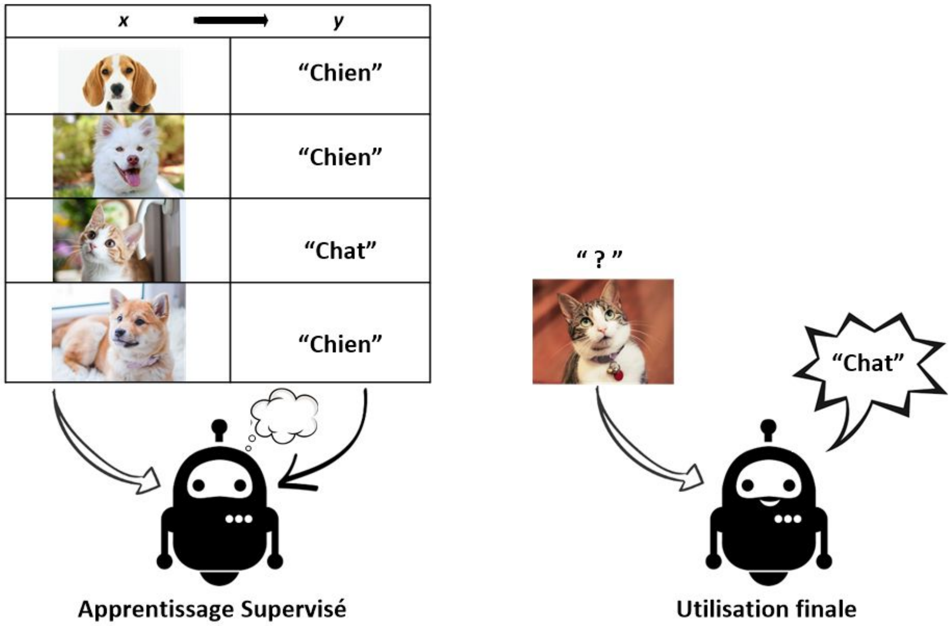
Quick Recap: Supervised Learning

Machine Learning Problem

Inputs :
- image - label pairs
- Initial model
- Metric to optimize

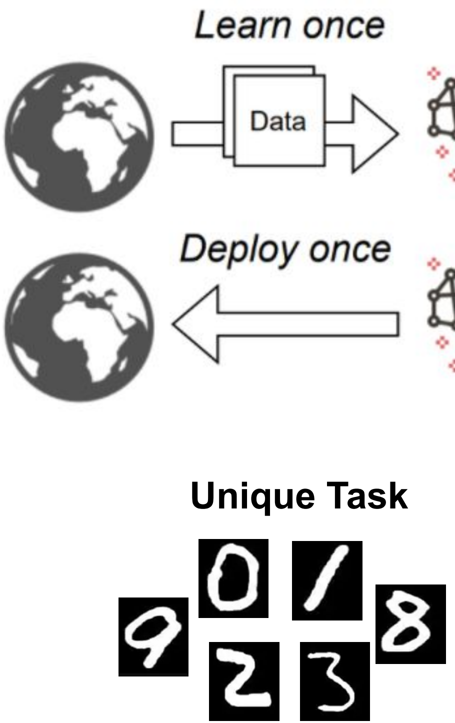
Outputs :
- trained model

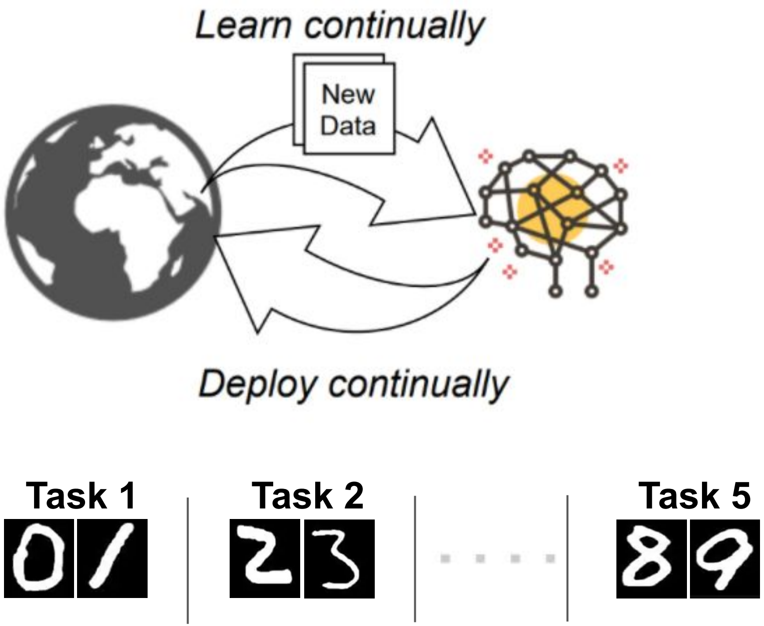
Objective : Generalize



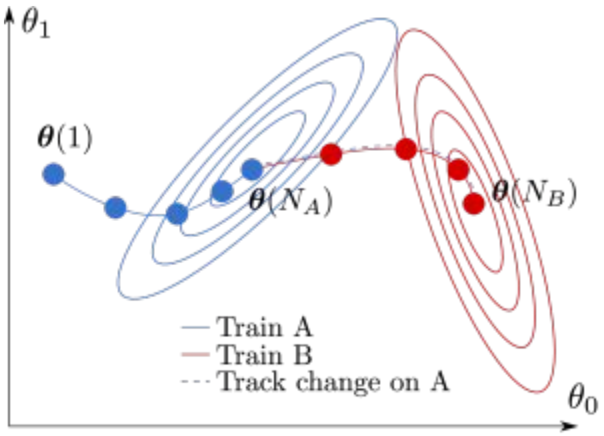
"Apprentissage Supervisé : Introduction," *Machine Learnia*, Jul. 02, 2019. https://machinelearnia.com/apprentissage-supervise-4-etapes/ (accessed Mar. 17, 2021).

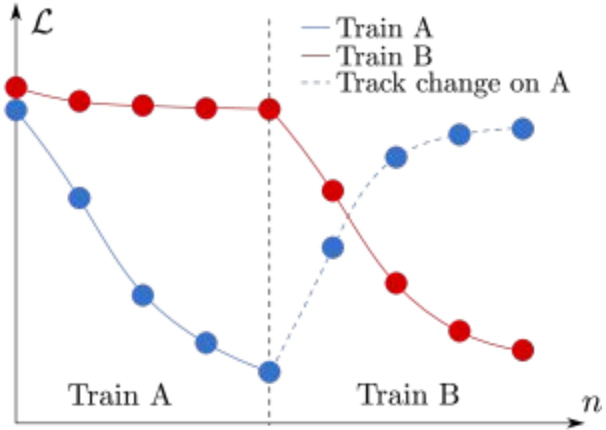
What is Continual Learning?





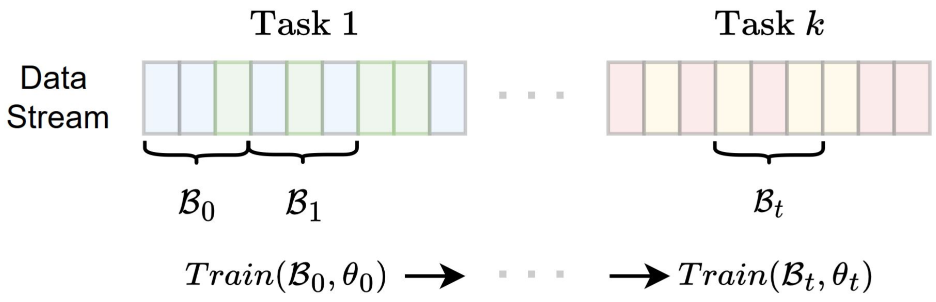
The problem: Catastrophic Forgetting





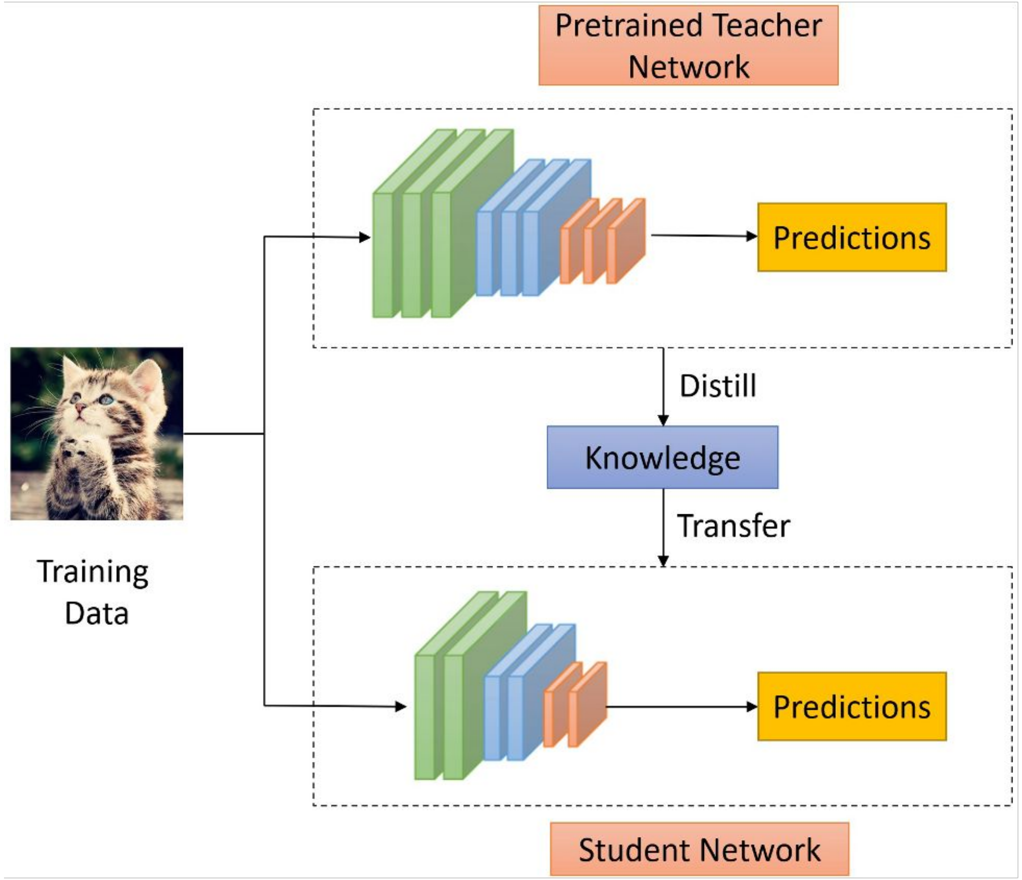
Method	Split-CIFAR10	Split-CIFAR100
offline	80.0	43.2
Incremental	16.4	3.6

Online CL VS Offline CL



Example: YouTube
Realistic: Can we store all YouTube data?

Knowledge Distillation



→ Small model mimics large model

Distillation challenges in OCL

Teacher Quality

Training Scenario	Accuracy (%)
Offline CL	81.8
Online CL	61.0
Online CL, Hard task	51.6
Online CL, Easy task	72.1

Method	Accuracy (%)
ER	49.0±4.6
ER+low qual. teach.	50.7±4.3
ER+high qual. teach.	54.6±3.3

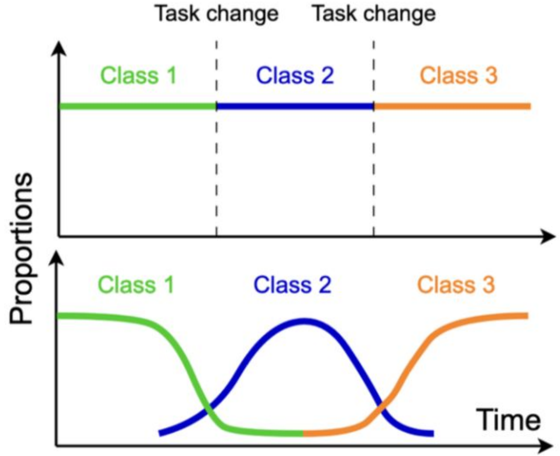
⚠ The quality of the teacher is not guaranteed (only one epoch)

⚠ Lower quality teacher can lead to little improvement

Teacher Quantity

⚠ One teacher per task is unrealistic (exploding memory cost)

Unknown Task Boundaries



⚠ When to take the snapshot for distillation?

Exponential Moving Average

$$\theta_\alpha(t) = \alpha * \theta(t) + (1 - \alpha) * \theta_\alpha(t-1)$$

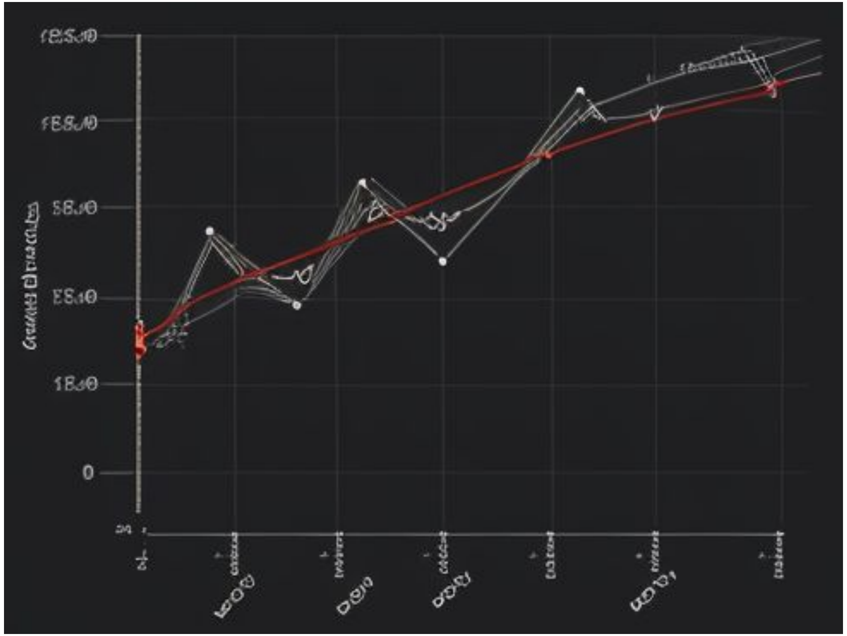
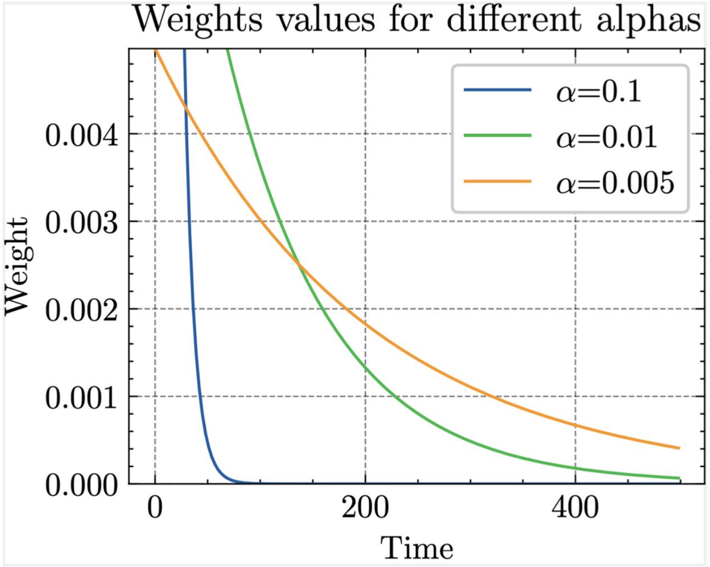
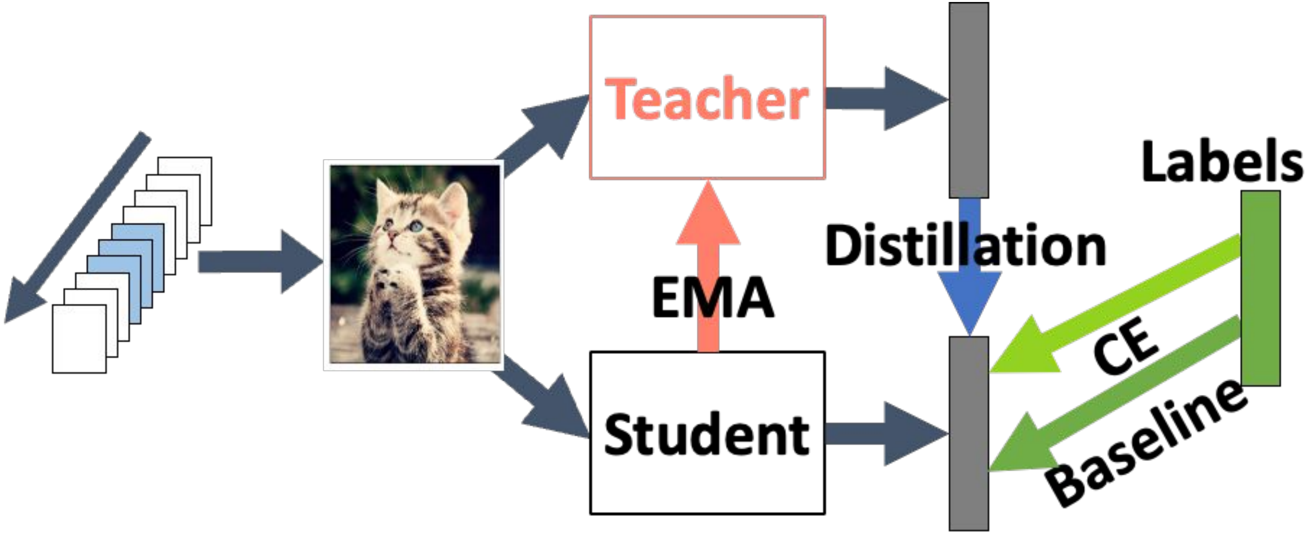


Image source: Me trying to generate some images



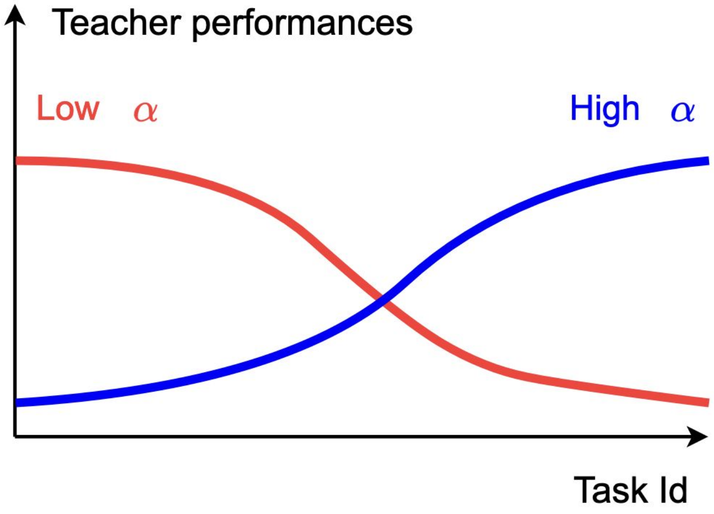
Momentum Knowledge Distillation



Stability-Plasticity control

Advantages

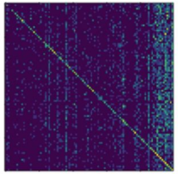
- Control of the teacher knowledge
- Only one teacher
- Evolving teacher
- No need for task boundaries



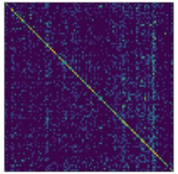
Results

Dataset	Tiny-IN		
Memory Size M	2000	5000	10000
ER [NeurIPS'19]	11.39 ±0.75	18.97 ±1.16	21.52 ±3.37
ER + SDP	15.32 ±0.47	23.22 ±0.31	26.97 ±1.1
ER + ours	**23.95** ±0.65	**32.22** ±0.88	**38.27** ±0.18
DER++ [NeurIPS'20]	3.89 ±0.64	4.28 ±0.51	4.16 ±0.32
DER++ + ours	**17.08** ±1.43	**15.64** ±4.64	**13.69** ±3.3
ERACE [ICLR'22]	14.79 ±0.95	22.25 ±1.69	26.64 ±0.91
ERACE + ours	**22.21** ±0.87	**31.13** ±0.41	**35.54** ±0.43
DVC [CVPR'22]	2.04 ±0.8	1.47 ±0.49	1.54 ±0.79
DVC + ours	**9.41** ±1.43	**12.03** ±3.83	**13.44** ±3.84
OCM [ICML'22]	19.58 ±0.63	27.85 ±1.03	32.56 ±1.37
OCM + ours	**23.07** ±0.37	**31.82** ±0.72	**37.46** ±0.95

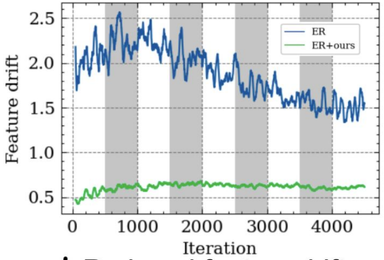
ER



ER + ours



Reduced task-recency bias



Method	CIFAR100	ImageNet100
ER	-16.7 ±1.2	-17.5 ±1.5
ER + ours	**+8.15 ±0.8**	**-1.3 ±2.3**
DER++	-27.5 ±3.4	-18.9 ±2.5
DER++ + ours	**-10.4 ±5.6**	**-14.4 ±2.5**
GSA	-4.9 ±1.2	-17.0 ±1.3
GSA + ours	**-2.5 ±3.1**	**-15.5 ±1.0**

Reduced feature drift

Improved backward transfer

- Solves many of OCL difficulties
- Small computation overhead
- Achieves a better stability-plasticity trade-off
- Simple yet efficient

Merci pour votre attention

Conclusions

- Offline and Online CL have different challenges
- Room for improvement in applying distillation in OCL

Presented papers

ICML'24

My co-authors

Giovanni Chierchia

Romain Negrel

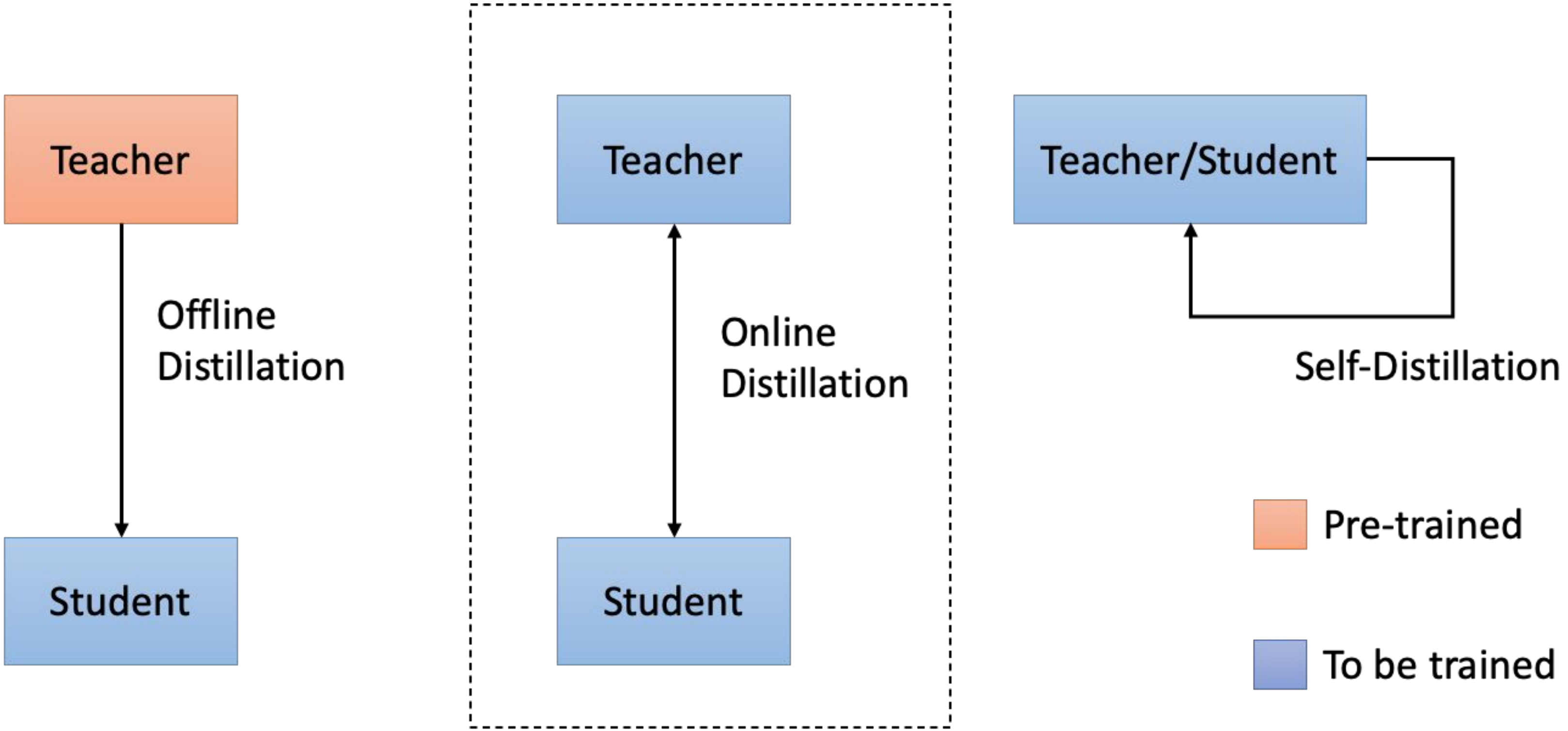
Jean-François Bercher

Toshihiko Yamasaki

Maorong Wang

Merci pour votre attention

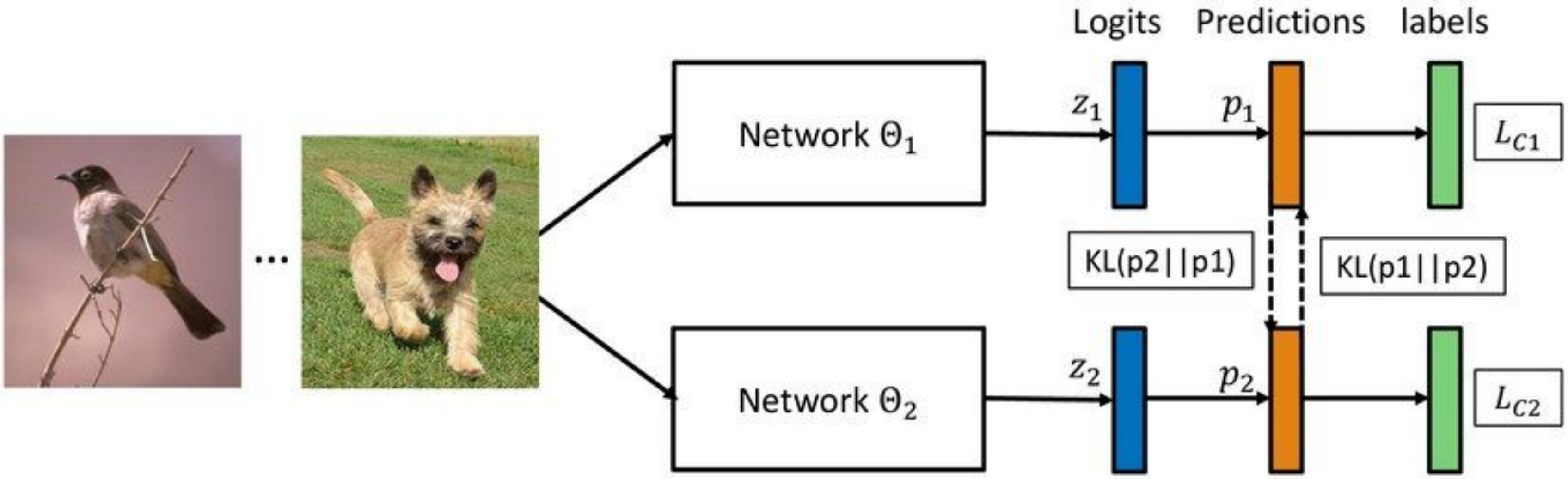
Knowledge Distillation Schemes



Three different knowledge distillation schemes

Improving plasticity with Collaborative Learning

Mutual Learning

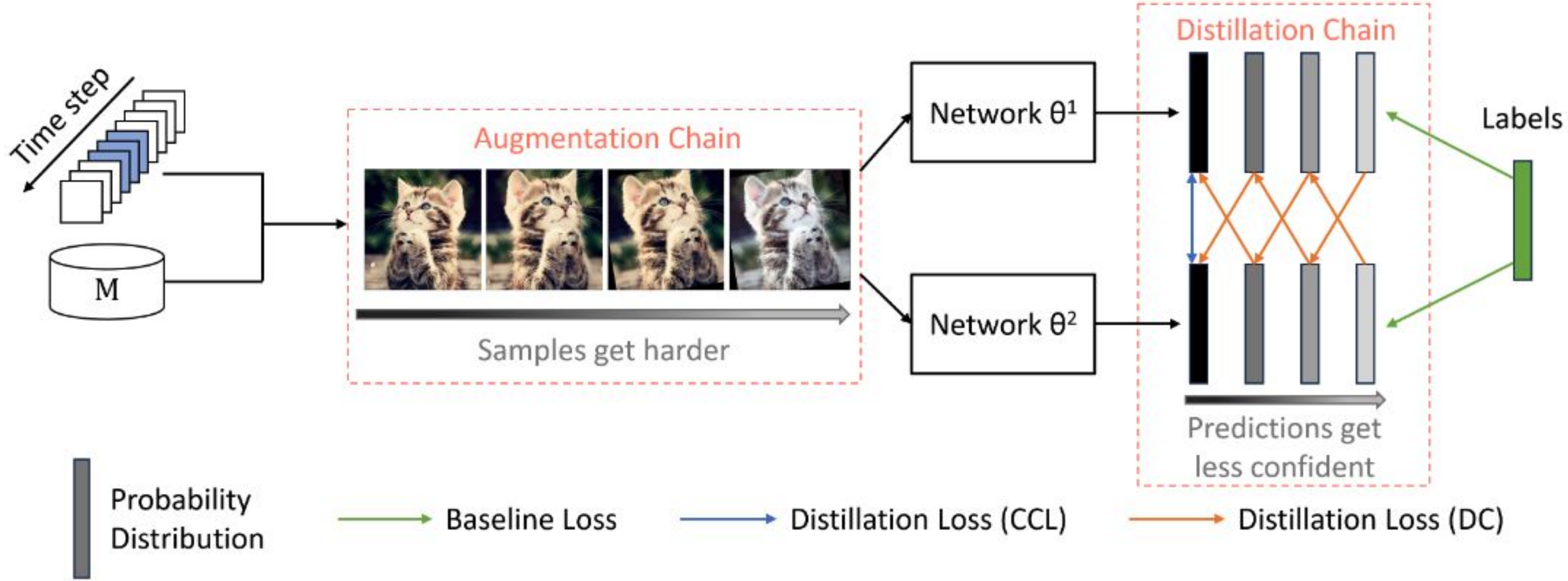


Why? Boost performance and convergence

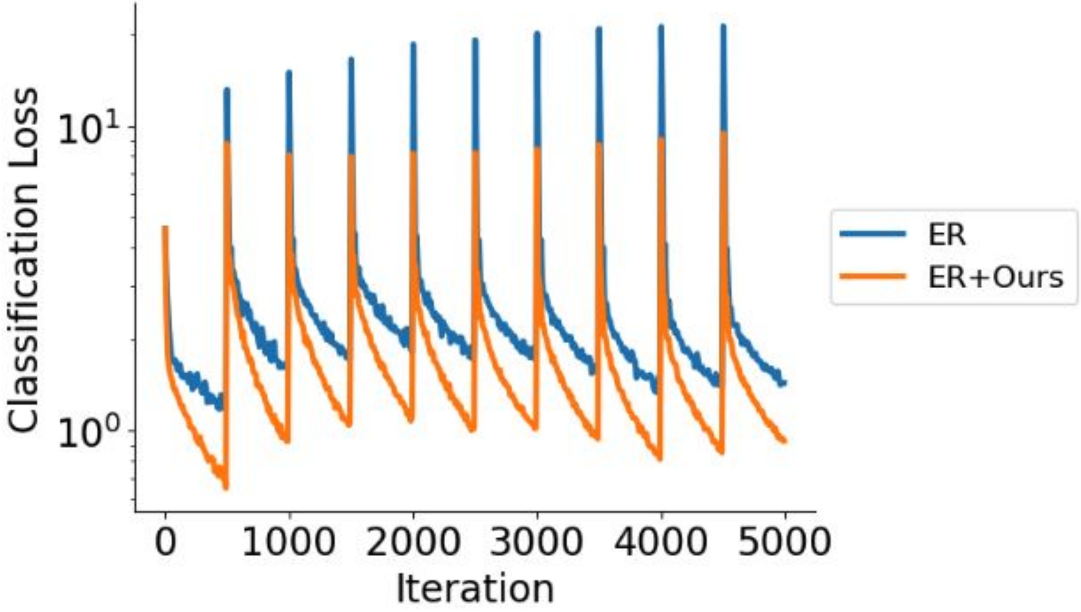
How? Randomness in the training process

source: A Selective Survey on Versatile Knowledge Distillation Paradigm for Neural Network Models, Ku et al. 2020.

Overall approach



Results



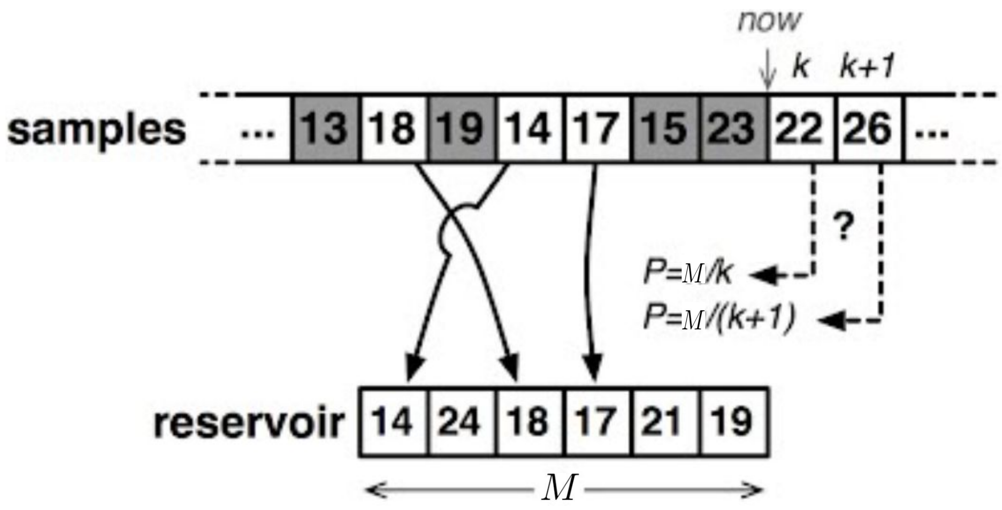
Classification Loss
10^1
10^0
0
1000
2000
3000
4000
5000
Iteration
ER
ER+Ours

Results

Dataset	CIFAR10		CIFAR100			Tiny-ImageNet			ImageNet-100
Memory Size M	500	1000	1000	2000	5000	2000	5000	10000	5000
ER [35]	56.68±1.89	62.32±4.13	24.47±0.72	31.89±1.45	39.41±1.81	10.82±0.79	19.16±1.42	24.71±2.52	33.30±1.74
ER + Ours	**66.43±2.48**	**74.10±1.71**	**33.43±1.06**	**44.45±1.04**	**53.81±1.16**	**16.56±1.63**	**29.39±1.23**	**37.73±0.85**	**43.11±1.49**
DER++ [6]	58.04±2.30	64.02±1.92	25.09±1.41	32.33±2.66	38.31±2.28	8.73±1.58	17.95±2.49	19.40±3.71	34.75±2.23
DER++ + Ours	**68.79±1.42**	**74.25±1.10**	**34.36±0.89**	**43.52±1.35**	**52.95±0.86**	**10.99±1.39**	**21.68±1.94**	**28.01±2.46**	**45.70±1.32**
ER-ACE [7]	53.26±3.04	59.94±2.40	28.36±1.99	34.21±1.53	39.39±1.31	13.56±1.00	20.84±0.43	25.92±1.07	38.37±1.20
ER-ACE + Ours	**70.08±1.38**	**75.56±1.14**	**37.20±1.15**	**45.14±1.00**	**53.92±0.48**	**18.32±1.49**	**26.22±2.01**	**32.23±1.70**	**45.15±1.94**
OCM [19]	68.19±1.75	73.15±1.05	28.02±0.74	35.69±1.36	42.22±1.06	18.36±0.95	26.74±1.02	31.94±1.19	23.67±2.36
OCM + Ours	**74.14±0.85**	**77.66±1.46**	**35.00±1.15**	**43.34±1.51**	**51.43±1.37**	**23.36±1.18**	**33.17±0.97**	**39.25±0.88**	**43.19±0.98**
GSA [20]	60.34±1.97	66.54±2.28	27.72±1.57	35.08±1.37	41.41±1.65	12.44±1.17	19.59±1.30	25.34±1.43	41.03±0.99
GSA + Ours	**68.91±1.68**	**75.78±1.16**	**35.56±1.39**	**44.74±1.32**	**55.39±1.09**	**16.70±1.66**	**28.11±1.70**	**37.13±1.75**	**44.28±1.16**
OnPro [44]	70.47±2.12	74.70±1.51	27.22±0.77	33.33±0.93	41.59±1.38	14.32±1.40	21.13±2.12	26.38±2.18	38.75±1.03
OnPro + Ours	**74.49±2.14**	**78.64±1.42**	**34.76±1.12**	**41.89±0.82**	**50.01±0.85**	**21.81±1.02**	**32.00±0.72**	**38.18±1.02**	**47.93±1.26**

- Small computation overhead (x2, but its ok)
- Achieves a better stability-plasticity trade-off

Réservoir Sampling



- Taille mémoire fixée
- Aucune information requise sur le stream
- Bonne représentation statistique dans la mémoire

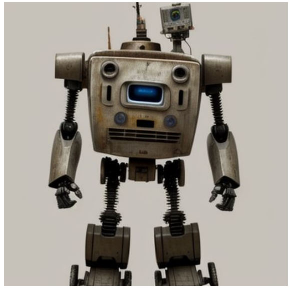
$$N_{updates}(K, M) = M\left(1 + \ln\frac{K}{M+1}\right)$$

[1]

J. S. Vitter, ‘Random sampling with a reservoir’, *ACM Trans. Math. Softw.*, vol. 11, no. 1, pp. 37–57, Mar. 1985, doi: 10.1145/3147.3165.

Stability–Plasticity trade-off

Intuitively

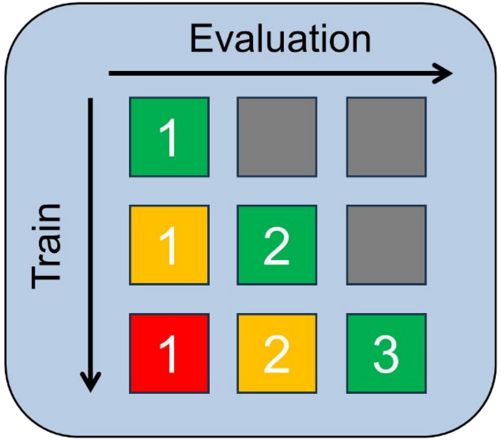


Stability: Retain old knowledge

Plasticity: Being able to acquire new knowledge

Image sources: Me trying to desperately use some generative AI in my slides

Some metrics



$$\text{Average Accuracy (AA)} = \frac{\boxed{1} + \boxed{2} + \boxed{3}}{3}$$

$$\text{Learning Accuracy (LA)} = \frac{\boxed{1} + \boxed{2} + \boxed{3}}{3}$$

$$\text{Forgetting Measure (FM)} = \frac{\left(\boxed{1} - \boxed{1}\right) + \left(\boxed{2} - \boxed{2}\right)}{2}$$

An example

$a_{k,j}$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	AA_k	AF_k
$\mathcal{T}_1$	50	-	-	-	-	50	0
$\mathcal{T}_2$	25	25	-	-	-	25	25
$\mathcal{T}_3$	16.7	16.7	16.7	-	-	16.7	20.83
$\mathcal{T}_4$	12.5	12.5	12.5	12.5	-	12.5	18.06
$\mathcal{T}_5$	10	10	10	10	10	10	16.04

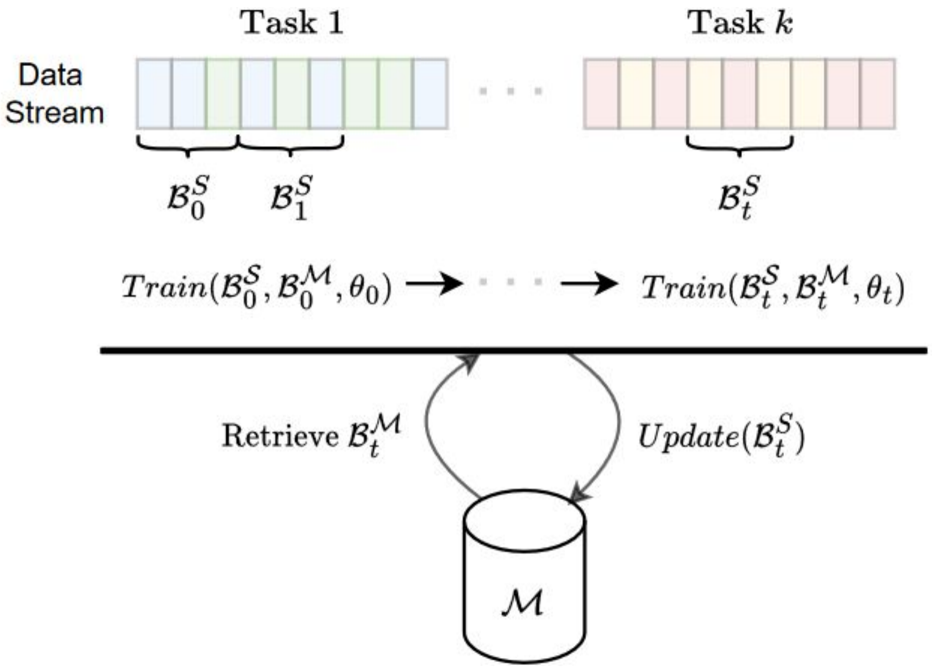
Plasticity in OCL

- In offline: main focus is stability, plasticity is not very challenging

- In online: plasticity is especially challenging

Why? -> One pass over the data is not enough

Back to memory based methods



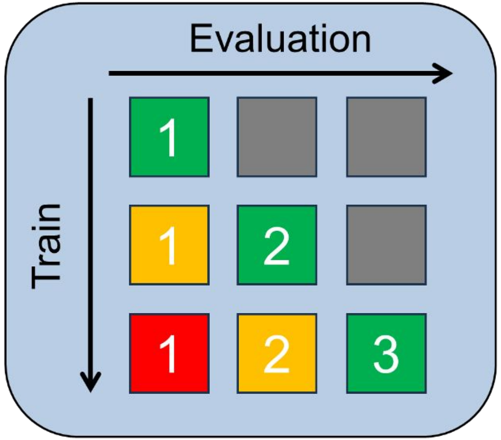
→ **Partially solves the lack of plasticity (multiple pass over memory data)**

→ **Can we do better?**

Another example

a_j^i	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	AA_k	LA_k	FM_k	RF_k
$\mathcal{T}_1$	30/15	-	-	-	-	30/15	30/15	-	-
$\mathcal{T}_2$	25/12.5	25/12.5	-	-	-	25/12.5	27.5/13.75	5/2.5	8.33/8.33
$\mathcal{T}_3$	20/10	20/10	20/10	-	-	20/10	25/12.5	7.5/3.75	17.78/17.78
$\mathcal{T}_4$	15/7.5	15/7.5	15/7.5	15/7.5	-	15/7.5	22.5/11.25	10/5	28.75/28.75
$\mathcal{T}_5$	10/5	10/5	10/5	10/5	10/5	10/5	20/10	12.5/6.25	42/42

More metrics



$$\text{Average Accuracy (AA)} = \frac{\boxed{1} + \boxed{2} + \boxed{3}}{3}$$

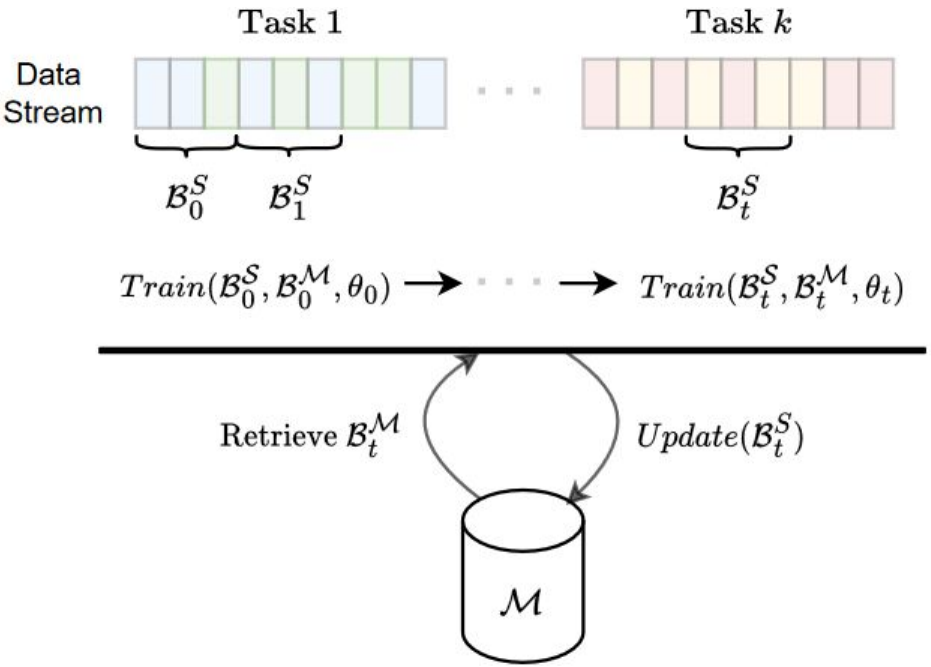
$$\text{Learning Accuracy (LA)} = \frac{\boxed{1} + \boxed{2} + \boxed{3}}{3}$$

$$\text{Forgetting Measure (FM)} = \frac{\left(\boxed{1} - \boxed{1}\right) + \left(\boxed{2} - \boxed{2}\right)}{2}$$

Proposed new metric

$$\text{Relative Forgetting (RF)} = \frac{\left(1 - \frac{\boxed{1}}{\boxed{1}}\right) + \left(1 - \frac{\boxed{2}}{\boxed{2}}\right) + \left(1 - \frac{\boxed{3}}{\boxed{3}}\right)}{3}$$

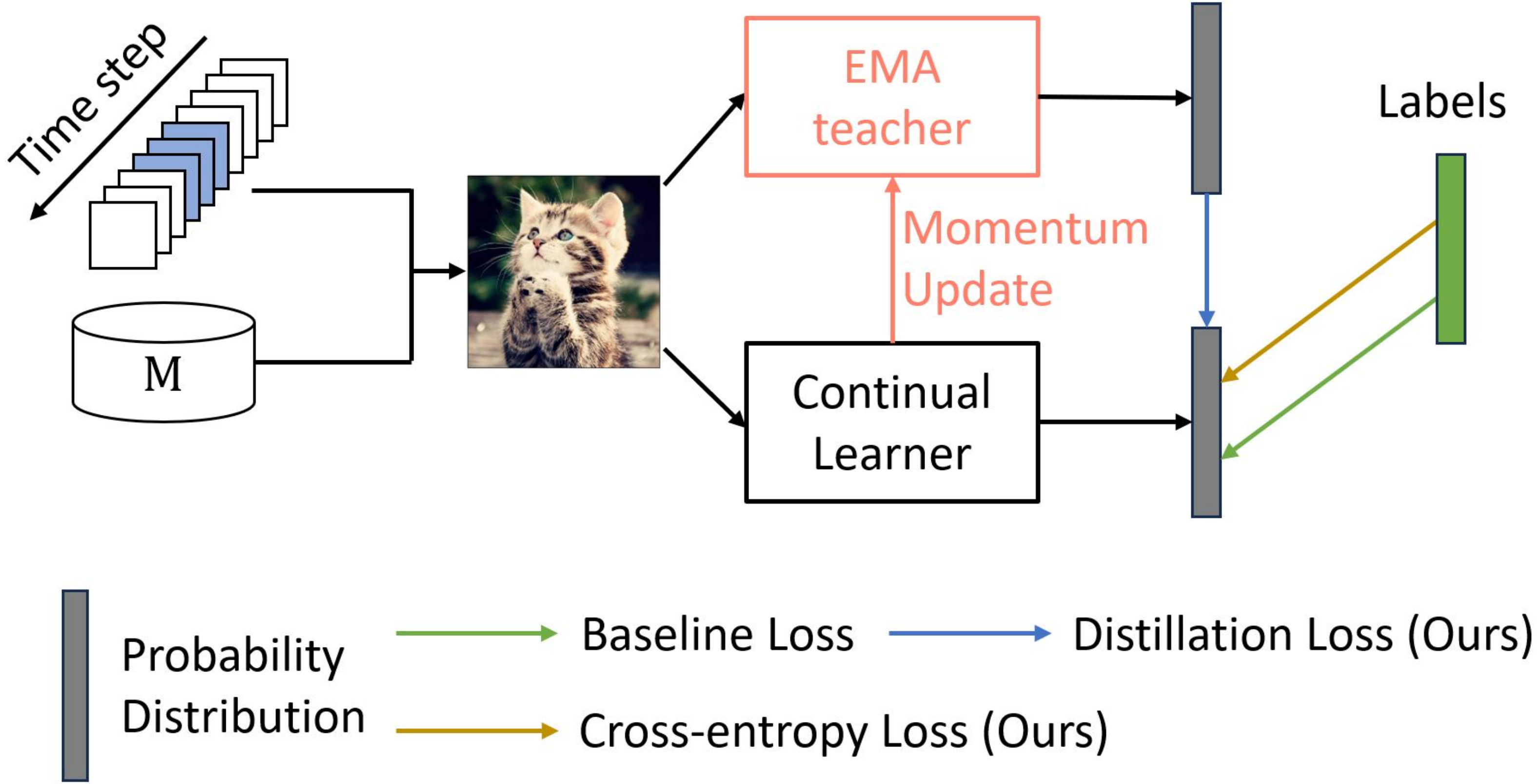
OCL State-of-the-art: Replay-based methods



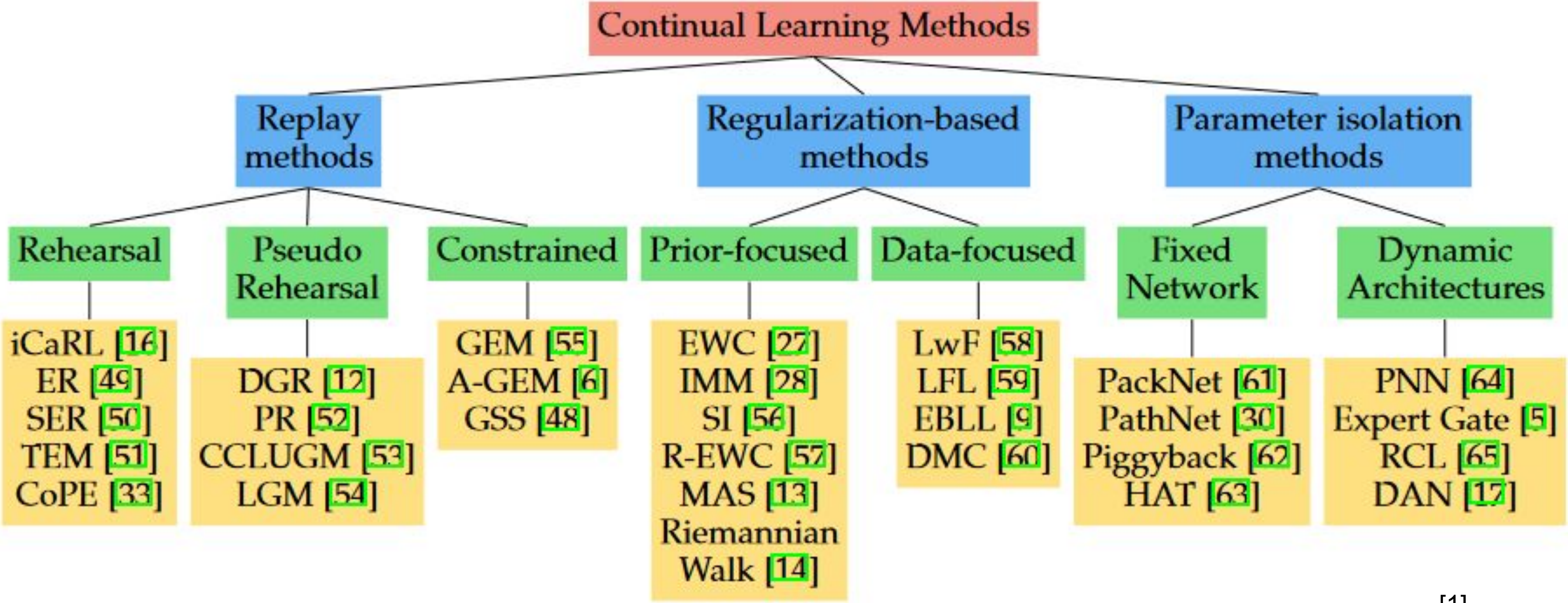
Generic Replay Algorithm

Input: Tasks $\{D_1, .., D_K\}$; Memory $\mathcal{M}$; Parameters θ
Output: Parameters θ; Memory $\mathcal{M}$
for $k \in \{1, .., K\}$ **do**
  for $\mathcal{B}_{\mathcal{S}} \in D_k$ **do**
    $\mathcal{B}_{\mathcal{M}} \leftarrow \ RandomRetrieval(\mathcal{M})$
    $\mathcal{B} \leftarrow \ \mathcal{B}_{\mathcal{S}} \cup \mathcal{B}_{\mathcal{M}}$
    $\theta \leftarrow \ TrainModel(\theta, \mathcal{B})$
    $\mathcal{M} \leftarrow \ ReservoirUpdate(\mathcal{B}_{\mathcal{S}}, \mathcal{M}, \theta(\text{optional}))$
return: θ; M

Momentum Knowledge Distillation



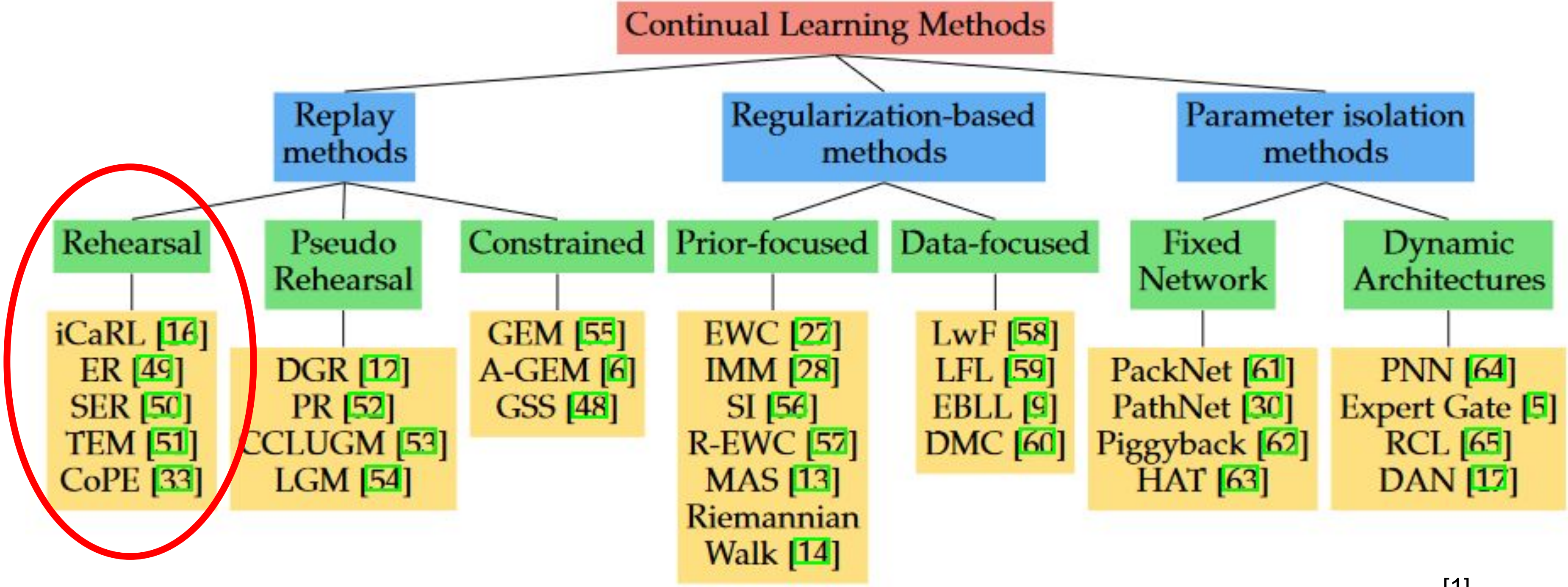
A lot of approaches



[1]

M. De Lange *et al.*, ‘A continual learning survey: Defying forgetting in classification tasks’, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3057446.

A lot of approaches



[1]

M. De Lange *et al.*, ‘A continual learning survey: Defying forgetting in classification tasks’, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3057446.