Rethinking Memory in Continual Learning: Beyond a Monolithic Store of the Past

Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet

AI Institute, University of Waikato

June, 2026

Background

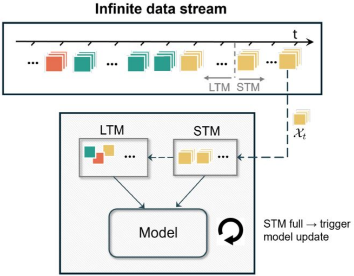
- Deep neural networks suffer from **Catastrophic Forgetting** when learning from non-IID data streams.
- The core problem in CL is balancing:
  - **Plasticity**: The ability to acquire new knowledge.
  - **Stability**: The ability to retain previous knowledge.
- Memory-based methods (storing past exemplars to replay) are one of the fundamental ideas to mitigate forgetting.

Memory in Continual Learning: A Dual-Memory System

- **Conventional View**:
  - CL memory is often treated as a monolithic store of past data
  - Online CL has worse empirical performance than offline CL
  - Online CL is **inherently more difficult** than offline CL
- **Our Insight**:
  - A Dual-Memory System with short-term and long-term memory
    - Analogy: Human memory systems also have a temporal buffer to hold the most recent experiences for a short period
  - Online and offline CL can be unified with the dual-memory replay framework
  - Online CL can lead to **comparable or better performance** than offline CL under the same compute budget

Unified Framework and Memory Capacity Ratio (MCR)

- A **Unified Framework** for online and offline CL



- **Memory Capacity Ratio (MCR)**: the distinction between the two

$$\lambda = \frac{\text{Capacity of } \mathcal{M}_{short}}{\text{Capacity of } \mathcal{M}_{long}}$$

  - **Offline CL**: large MCR, M_short = task size
  - **Online CL**: small MCR, M_short = batch size

Theoretical Study on Generalization Bound with 0-1 Loss

- **Theorem 1** provides a generalization bound for h:

$$\mathcal{L}_{\mathcal{D}_t}(h, h^*_y) \leq \hat{\mathcal{L}}_{\mathcal{M}_t}(h, h^*_{M_t}) + \hat{\Re}_{\mathcal{M}_t}(H) + +\mathcal{L}_{\mathcal{D}_t}(h^*_{M_t}, h^*_{D_t}) + 3A_0\sqrt{\frac{\log \frac{2}{\delta}}{2M}} + \mathrm{disc}_L(\mathcal{D}_t, \mathcal{M}_t) + \mathcal{L}_{\mathcal{D}_t}(h^*_{M_t}, h^*_y)$$

- **Proposition 1** (Relating MCR to disc_L):

$$\mathrm{disc}_L(\mathcal{D}_t, \mathcal{M}_t) = \frac{N_t^- \lambda(N_t - M)}{N_t(N_t + \lambda N_t - \lambda M)} \mathrm{disc}_L(\mathcal{D}_t^-, \mathcal{D}_t^+)$$

- **Corollary 1** provides insights on the gradient w.r.t. λ (MCR):

$$\nabla_\lambda \mathrm{disc}_L = \frac{N_t(N_t - M)}{(N_t + \lambda N_t - \lambda M)^2} \mathrm{disc}_L(\mathcal{D}_t^-, \mathcal{D}_t^+) \geq 0$$

Theoretical analysis in a simplified setting shows that a small MCR leads to a lower generalization bound.

Empirical Study: the effect of MCR

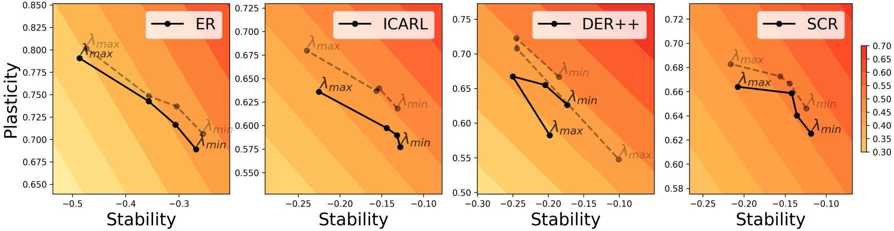
- **Control variables**:
  - Stored Data Size (M)
  - Seen Data Size (N)
  - Compute Resources (K)
  - Algorithms and Hyperparameters
- **Forgetting mitigation strategies**
  - e.g., regularization loss, contrastive loss

	MCR	ER	iCARL	DER++	SCR
S-CIFAR100-20	λ_{max} (offline CL)	35.1 ± 0.8	44.8 ± 1.1	47.2 ± 0.4	45.1 ± 0.4
	λ_{min} (online CL)	$\mathbf{44.6 \pm 2.4}$	$\mathbf{50.0 \pm 1.6}$	$\mathbf{50.9 \pm 0.9}$	$\mathbf{51.9 \pm 0.5}$
	Δ	$9.5 \uparrow$	$5.2 \uparrow$	$3.7 \uparrow$	$6.8 \uparrow$
S-MINI-IMAGENET-10	λ_{max} (offline CL)	31.2 ± 1.1	42.5 ± 1.2	41.0 ± 2.7	46.3 ± 0.4
	λ_{min} (online CL)	$\mathbf{43.7 \pm 1.2}$	$\mathbf{46.8 \pm 1.5}$	$\mathbf{46.4 \pm 1.3}$	$\mathbf{51.8 \pm 0.7}$
	Δ	$12.5 \uparrow$	$3.3 \uparrow$	$5.4 \uparrow$	$5.5 \uparrow$
S-CORE-9	λ_{max} (offline CL)	40.1 ± 2.4	45.8 ± 1.6	38.2 ± 1.5	62.1 ± 1.3
	λ_{min} (online CL)	$\mathbf{50.7 \pm 1.7}$	$\mathbf{50.1 \pm 1.6}$	$\mathbf{46.7 \pm 2.6}$	$\mathbf{69.7 \pm 0.3}$
	Δ	$10.6 \uparrow$	$4.3 \uparrow$	$8.5 \uparrow$	$7.6 \uparrow$

Table: Performance Comparison across different MCRs

Empirical Study: Stability, Plasticity, and Generalization

- **The Effect of Reducing MCR**:
  - Yields diverse stability–plasticity patterns.
  - **Consistently improves the overall stability–plasticity trade-off**.



Solid/dashed lines: two seeds (varied initialization / task split)

Conclusion

- **A New Perspective**: There is a **dual-memory system** that arises in both online and offline CL settings.
  - **Online CL** maintains a small short-term memory (a **small MCR**).
  - **Offline CL** maintains a large short-term memory (a **large MCR**).
- **Main Findings**:
  - Theoretical analysis in a simplified setting shows small MCR leads to a lower generalization bound
  - Empirically, reducing the MCR consistently matches or outperforms larger MCRs in generalization.
- **Take-away Message**: The **undervalued potential of online CL**.
  - For budgeted resources, online /semi-online CL matches or outperforms offline CL

Thank you!

- Email: yaqianz@waikato.ac.nz
- TAIAO: https://taiao.ai/
- AI Institute: https://ai.waikato.ac.nz/