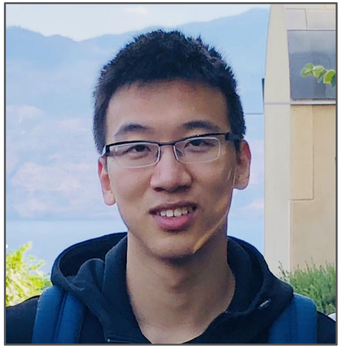
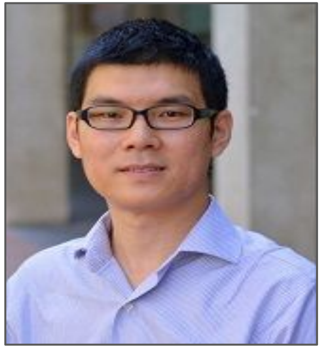
Navigating Memory Construction by Global Pseudo-Task Simulation for Continual Learning

(Paper Id: 6620)

Yejia Liu*
UC Riverside



Wang Zhu*
USC

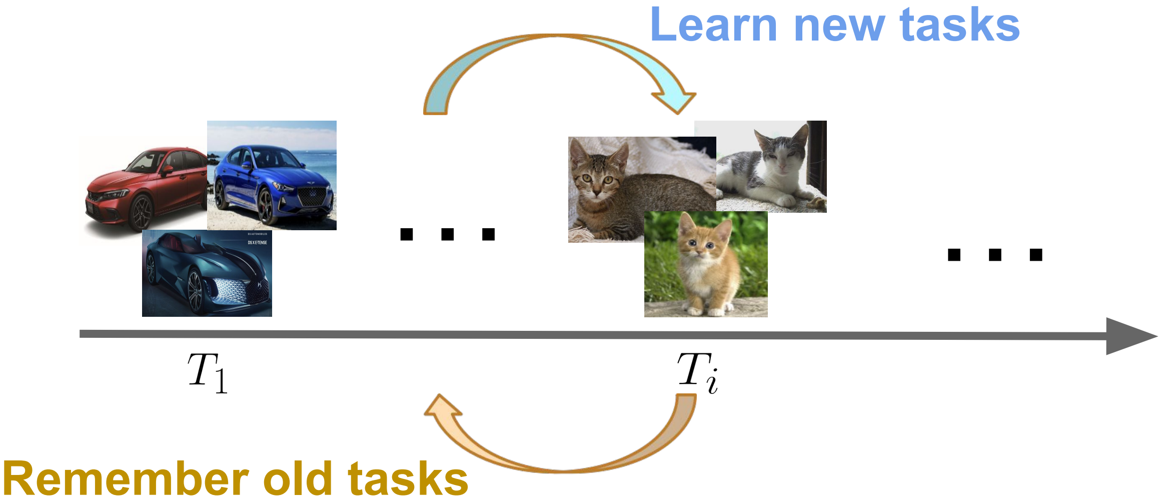


Shaolei Ren
UC Riverside

(: authors contributed equally)*

Continual Learning (CL)

- Desiderata
  - Adapt new knowledge
  - Maintain old skills
  - Constant memory
  - Zero-shot learning
  - Backward transfer



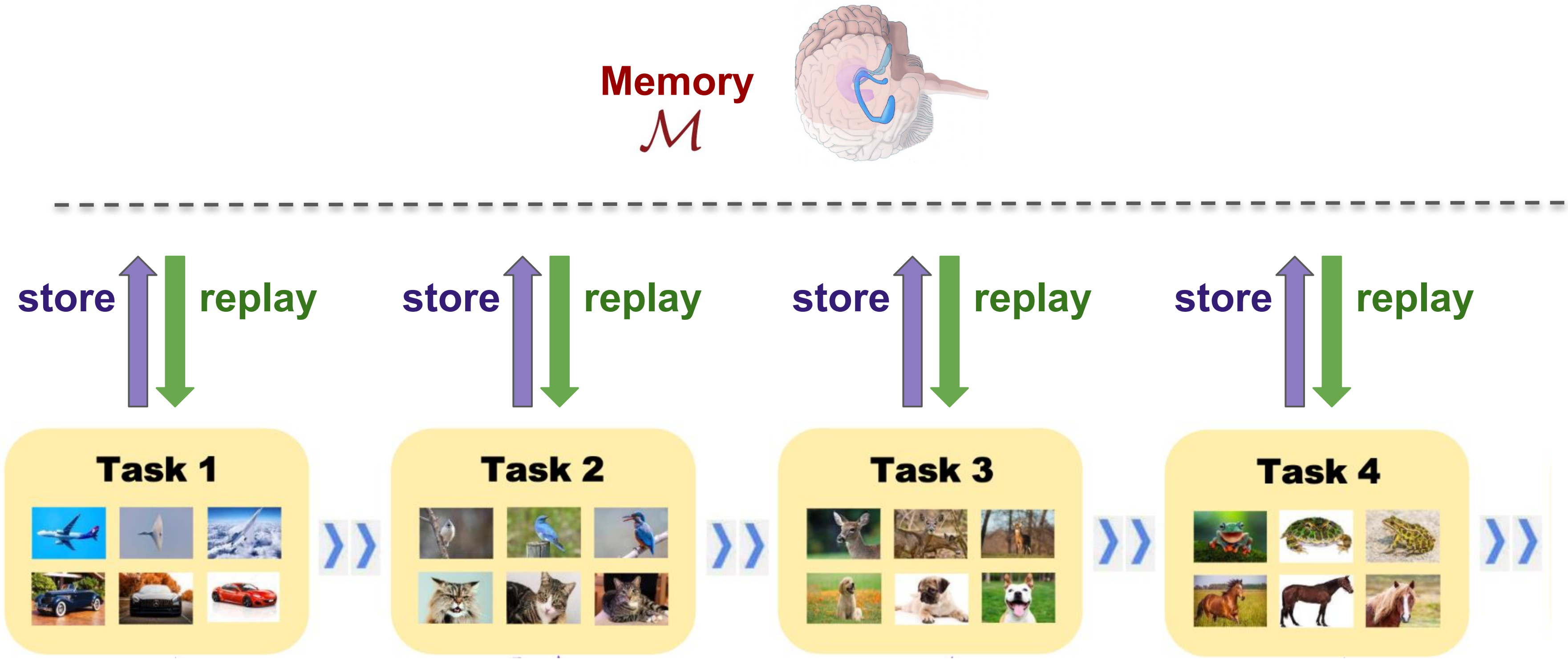
$$\boxed{\mathcal{L}^G} = \sum_{i=1}^{T} \mathbb{E}_{(x_i, y_i) \sim P_i} \ell(y_i, f(x_i; \boxed{\theta_T}))$$

global loss

final model parameters

(Image credit: CIFAR10)

Experience Replay (ER)



Challenges

Challenge 1: global objective - L^G has no direct relationship with $\mathcal{M}$

- Most ER works optimize one single task with the memory buffer at a time
- The objective of CL is to achieve the minimum loss across *all* experienced tasks

Challenge 2: long task sequence - a remaining challenge in the current CL regime

- Random sampling squeezes out class representations
- Per-class sampling not exploits the generalization benefits of random sampling
- **Lack of dynamic memory construction policy**

Inspiration – Global Pseudo-Task Simulation (GPS)

- Explicitly optimize the global loss as a function of memory configuration

global loss

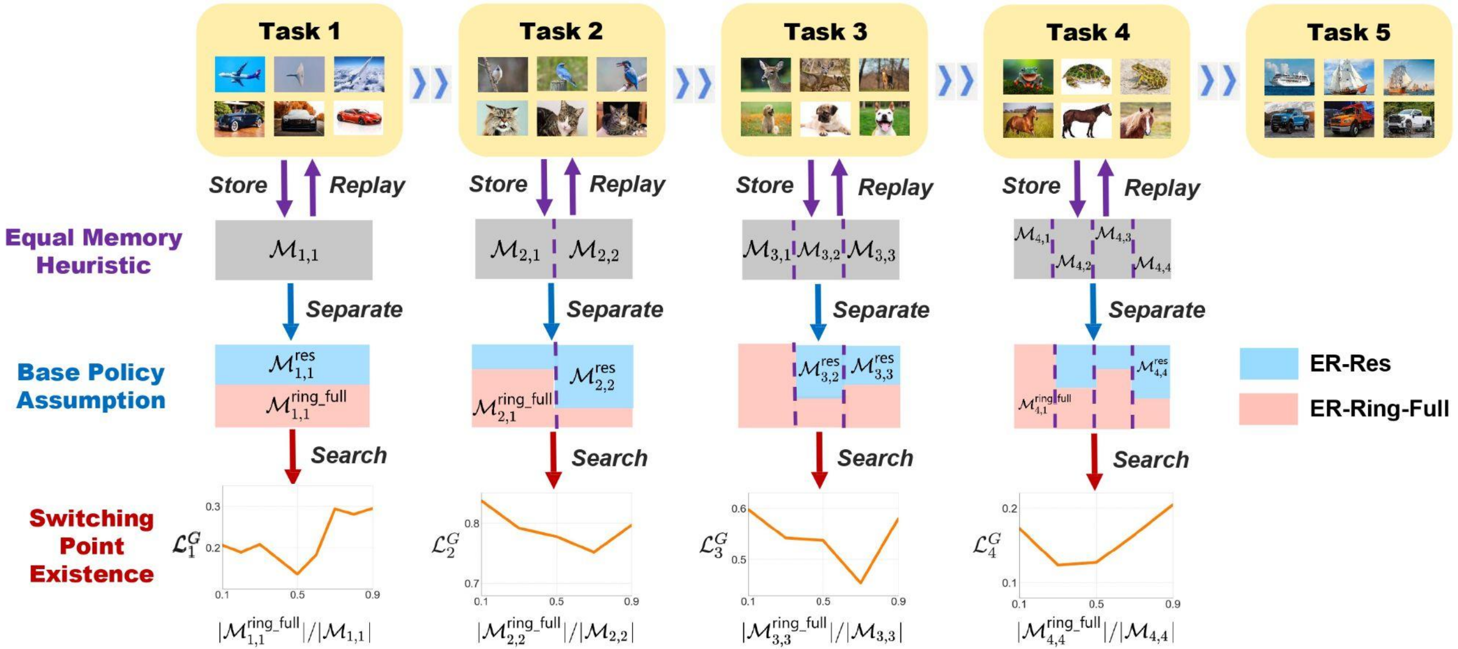
$$\min_{\{\mathcal{M}_i\}_{i\in\mathcal{T}}} \boxed{\mathcal{L}^G}(\{\boxed{\mathcal{M}_i}\}_{i\in\mathcal{T}})$$

memory after training task t_i

- Mimic the forgetting pattern for current task caused by future tasks
    - Similar learning difficulty of individual task on widely used vision benchmarks
    - Limited zero-shot transfer ability

A Starting Point: Offline Setup

- Res and Ring-Full mixed policy based on the switching point for each task.



Global Pseudo-task Simulation (GPS): Online Setup

- Approximate the switching point

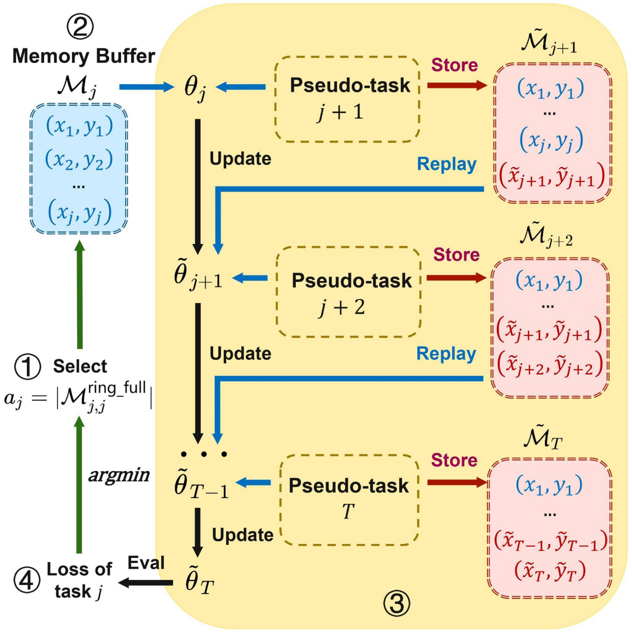
approx switching point

simulated final parameters

$$\boxed{\tilde{s}_j} = \arg\min_{\boxed{a_j}} \mathbb{E}_{(x_j,y_j)\sim P_j} \ell(y_j, f(x_j; \boxed{\tilde{\theta}_{j:T}}))$$

ring size of *task* t_i

  - Objective function containing simulation
  - Synthesize pseudo task with similar difficulty
  - Pseudo memory construction by binary search



Key Results: Compare to ER Baselines

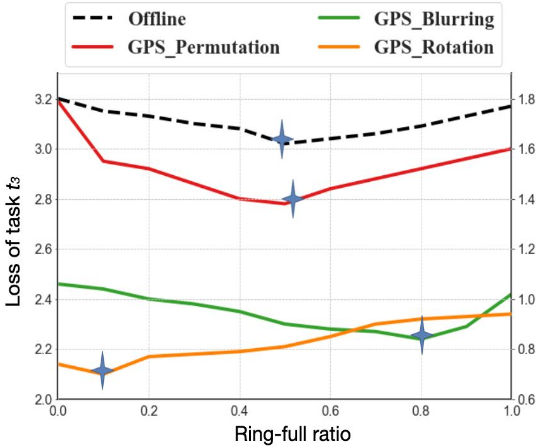
- Dynamic memory construction outperforms ER baselines and provide stability
- GPS using permutation is a good approximation

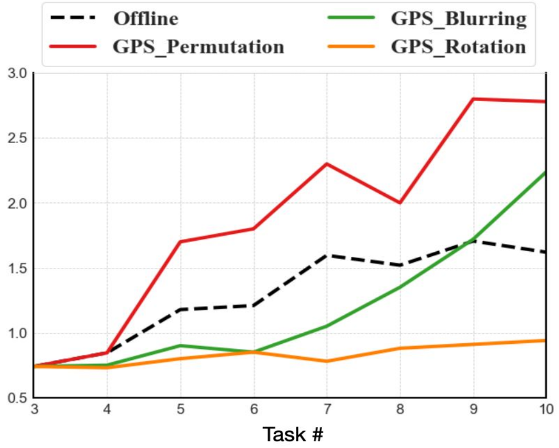
Method $\|\mathcal{M}\|$	Simulation	**P-MNIST** 1000	**S-CIFAR-10** 200	**S-CIFAR-100** 2000	**TinyImageNet** 2000
ER-Res	-	$86.55_{\pm 0.48}$	$92.01_{\pm 0.80}$	$81.38_{\pm 0.51}$	$57.50_{\pm 0.54}$
ER-Ring-Full	-	$84.33_{\pm 0.65}$	$91.53_{\pm 0.56}$	$81.16_{\pm 0.65}$	$54.73_{\pm 0.32}$
ER-Hybrid	-	$86.84_{\pm 0.35}$	$92.06_{\pm 0.89}$	$81.47_{\pm 0.23}$	$57.97_{\pm 0.44}$
GPS	**Permutation**	$\mathbf{87.93}_{\pm 0.21}$	$\mathbf{92.77}_{\pm 0.39}$	$\mathbf{82.46}_{\pm 0.33}$	$\mathbf{59.26}_{\pm 0.31}$
	Rotation	$85.38_{\pm 0.20}$	$91.61_{\pm 0.49}$	$81.50_{\pm 0.42}$	$57.45_{\pm 0.33}$
	Blurring	$86.03_{\pm 0.31}$	$91.96_{\pm 0.38}$	$81.49_{\pm 0.46}$	$56.85_{\pm 0.27}$
ER-Oracle	Offline	$88.26_{\pm 0.15}$	$93.09_{\pm 0.35}$	$82.88_{\pm 0.31}$	$60.56_{\pm 0.23}$

Analysis: Simulation Methods

Compared to offline:

- Rotation: too good zero-shot transfer
- Blurring: growing task difficulty
- **Permutation**: bear the **closest** switching point





S-CIFAR-100

Analysis: Long Task Sequence

- GPS outperforms baselines even more on longer task sequences.
- Longer task sequence requires more careful memory construction

Method	$T = 20$	$T = 40$
ER-Res	$71.25_{\pm 1.07}$	$47.33_{\pm 2.75}$
ER-Ring-Full	$71.92_{\pm 1.47}$	$49.26_{\pm 2.66}$
ER-Hybrid	$72.27_{\pm 0.88}$	$49.63_{\pm 2.61}$
GPS	$\mathbf{74.63_{\pm 0.20}}$	$\mathbf{53.57_{\pm 0.63}}$
ER-Oracle	$75.21_{\pm 0.17}$	$54.44_{\pm 0.51}$

Extend P-MNIST to longer task sequence

Analysis: Integrate with Advanced ER variants

- GPS integrates well with existing advanced ER variants
- GPS outperforms other non-ER methods

	P-MNIST	TinyImageNet
oEWC	$69.21_{\pm 2.92}$	$20.81_{\pm 0.95}$
iCaRL	-	$38.77_{\pm 3.68}$
GSS	$86.34_{\pm 4.28}$	-
A-GEM	$77.36_{\pm 1.28}$	$25.30_{\pm 0.87}$
OGD	$81.52_{\pm 2.21}$	-
HAL	$87.69_{\pm 0.34}$	-
GPS+HAL	$\mathbf{88.23}_{\pm 0.03}$	-
DER++	$91.14_{\pm 0.22}$	$60.67_{\pm 1.08}$
GPS+DER++	$\mathbf{91.64}_{\pm 0.16}$	$\mathbf{61.01}_{\pm 0.98}$

Takeaways

Summary

- We explicitly formulate the dynamic memory construction of continual learning w.r.t. the global loss
- Simulation by permutation well approximates the offline switching point
- GPS performs well in the long task sequence of continual learning

Limitation

- Setting: focus on task- and domain-incremental
    - Extend to the class-incremental in the future
- New dataset benchmarks closer to the real world is needed
    - Task sequences having specific zero-shot patterns

arXiv: https://arxiv.org/abs/2210.08442
Github: https://github.com/liuyejia/gps_cl