USC University of Southern California

USC Viterbi School of Engineering

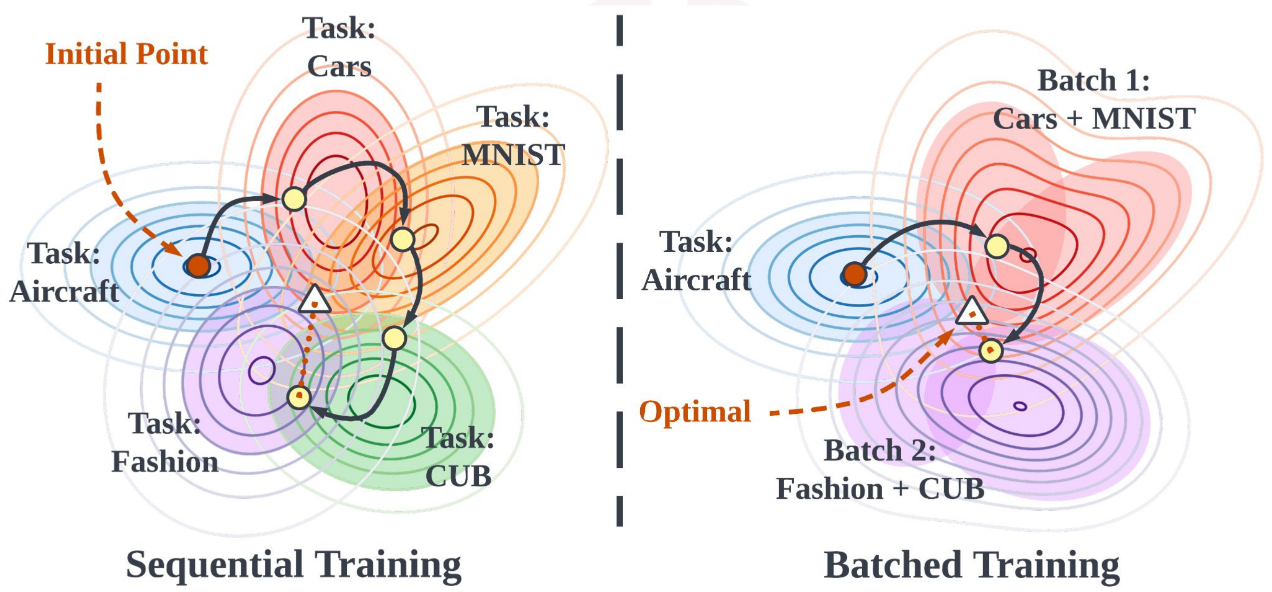
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Batch Model Consolidation: A Multi-Task Model Consolidation Framework

Iordanis Fostiropoulos, Jiaye Zhu, Laurent Itti

Overview

Previous approaches in **Continual Learning** suffer significant performance degradation and unacceptable cost when faced with a large number of diverse tasks [1, 2].

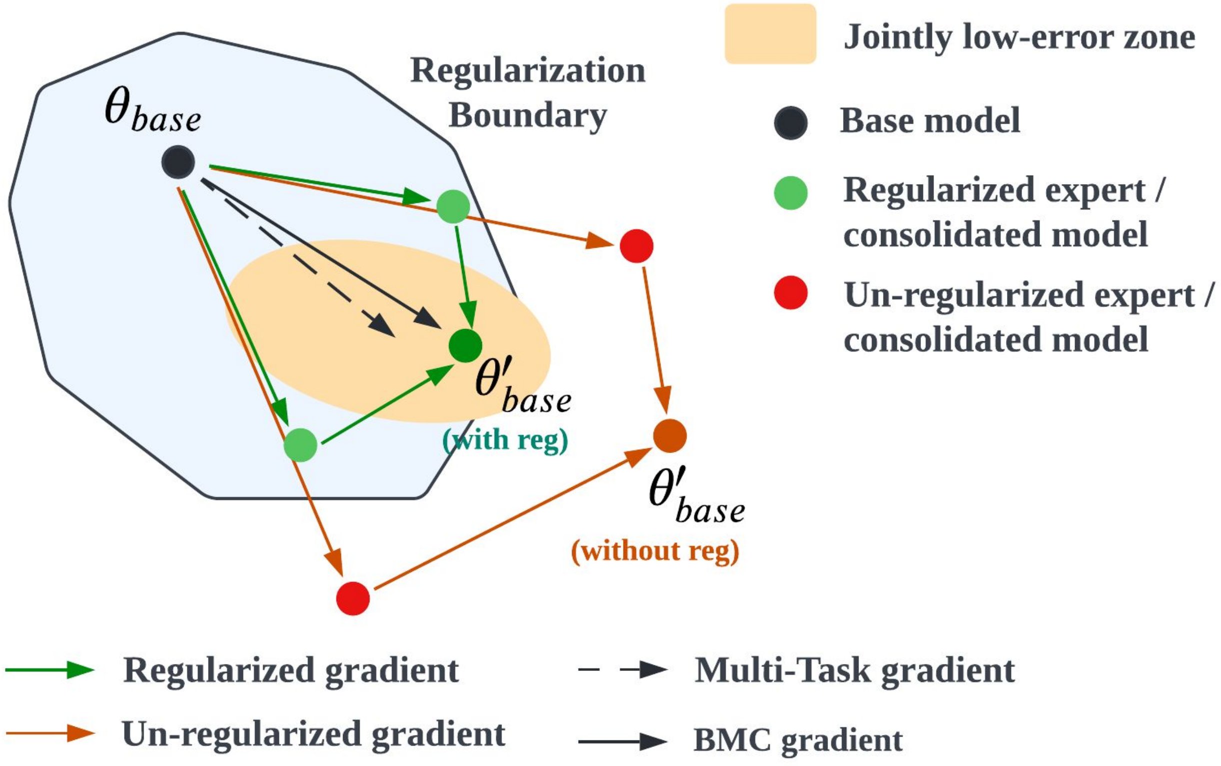


Our contributions:

- Propose **Batch Model Consolidation (BMC)** and a **distributed learning framework** to support CL for training multiple expert models on a single task stream composed of tasks from diverse domains.
- Propose a **stability loss** as regularization to expert models and a **batched distillation loss** combines multiple expert models to update a single base model in a single incremental step.
- We introduce **Stream dataset of 71 image classification tasks** and show that BMC is robust against large domain-shifts and for a large number of tasks.

USC

BMC - The Intuition

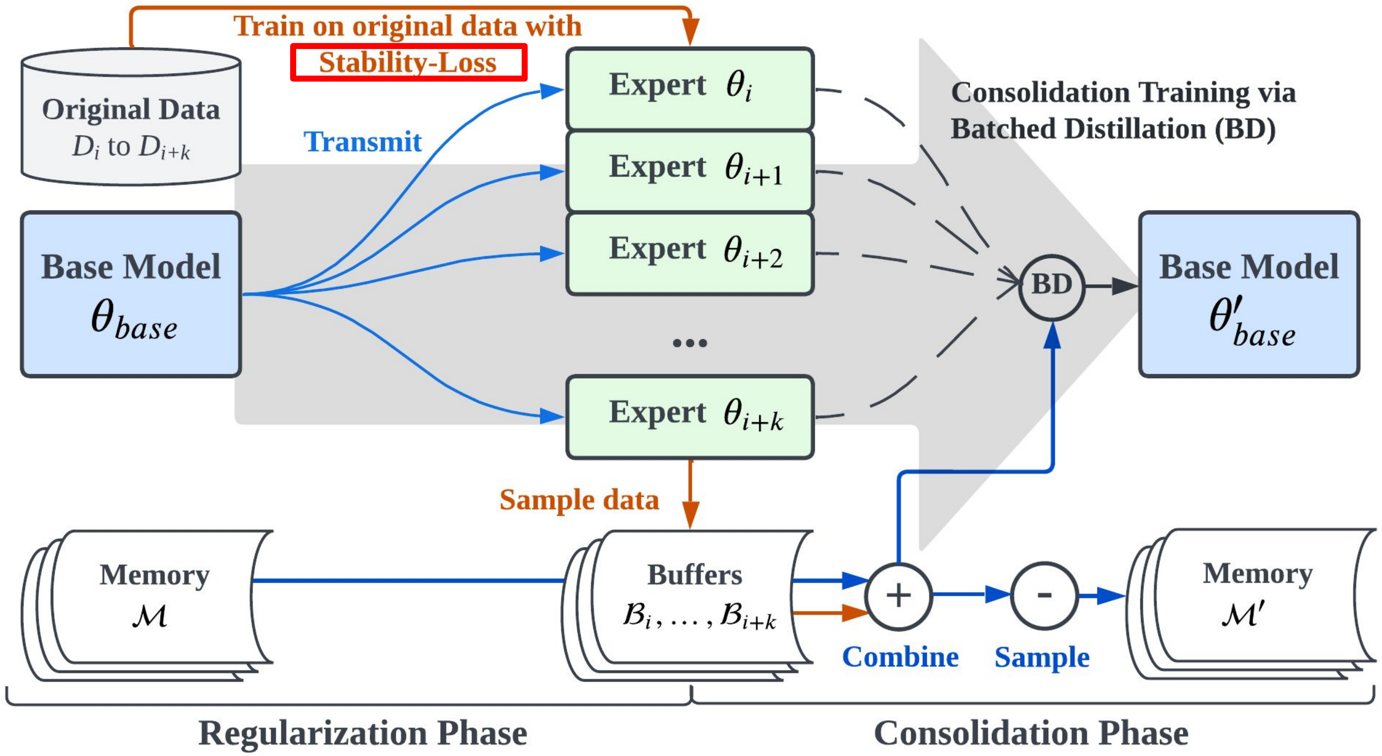


- We train the base model θ_{base} incrementally by regularized experts ●● to get the new base model θ'_{base}.

- Batched consolidation reduces gradient noise from distant tasks, and regularization improves the stability of base model.

BMC - Regularization Phase

A single incremental step of BMC



Interim. Feature Knowledge Distillation

Applied between experts and base model

$$\boxed{\mathcal{L}_{bd}(\theta_t(x), \theta_s(x))} = \sum_{i=1}^{|\theta|} ||sg(\phi_i^t) - \phi_i^s||_2$$

Stability Loss - Regularizing experts

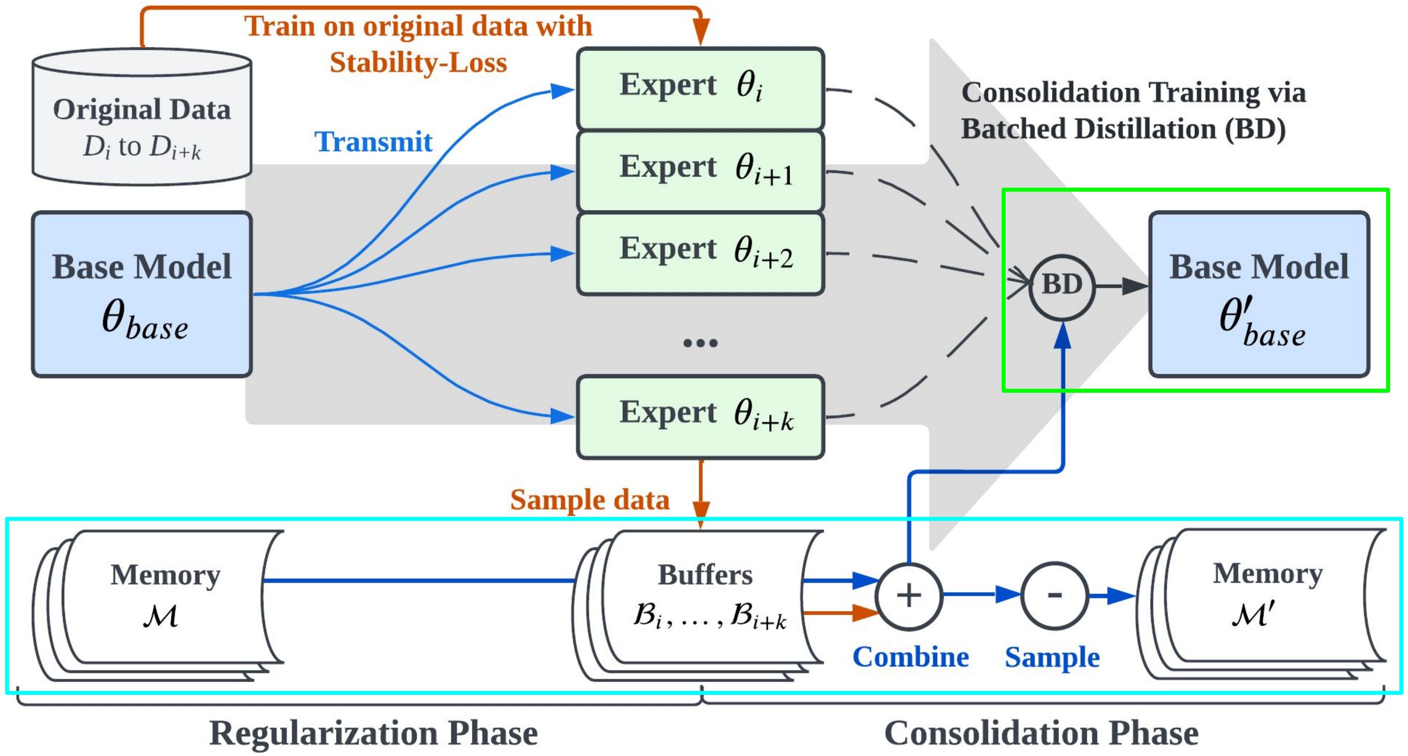
Used on the expert device and between the base model

$$\mathcal{L}_{exp} = \mathcal{L}_T(\theta_{exp}(x), y) + \boxed{\lambda \mathcal{L}_{bd}(\theta_{base}(x), \theta_{exp}(x))}$$

After experts training: sample consolidation artifacts as *Buffers*

BMC - Consolidation Phase

A single incremental step of BMC



Interim. Feature Knowledge Distillation

Applied between experts and base model

$$\boxed{\mathcal{L}_{bd}(\theta_t(x), \theta_s(x))} = \sum_{i=1}^{|\theta|} ||sg(\phi_i^t) - \phi_i^s||_2$$

Batched Distillation Loss

Apply distillation between multiple experts $\mathcal{E}$ and a single base model on data

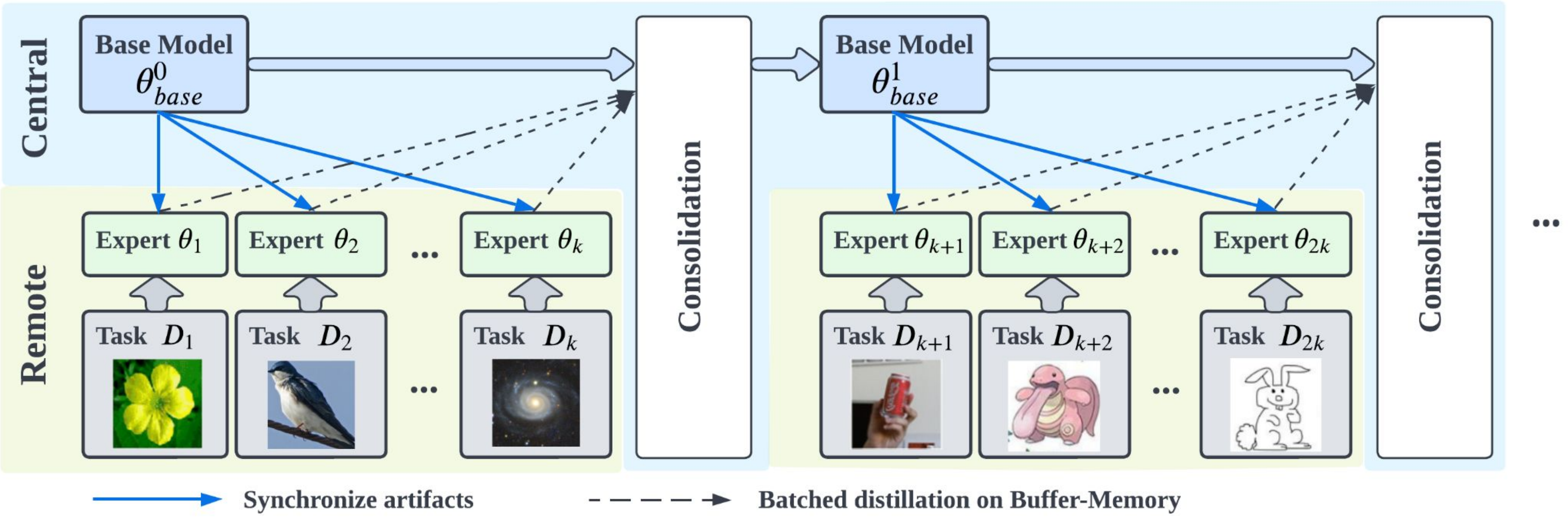
$$\boxed{\mathcal{D} = \{\mathcal{M}, \mathcal{B}_i, \ldots, \mathcal{B}_{i+k}\}}$$

$$\boxed{\mathcal{L}_{bmc}} = \sum_{\theta_i \in \mathcal{E}} \mathbb{E}_{x,\phi(x;\theta_i)\sim\mathcal{D}} \boxed{[\mathcal{L}_{bd}(\theta'_{base}(x), \theta_i(x))]}$$

Task Loss and Batched Distillation Loss is applied on θ'_{base}

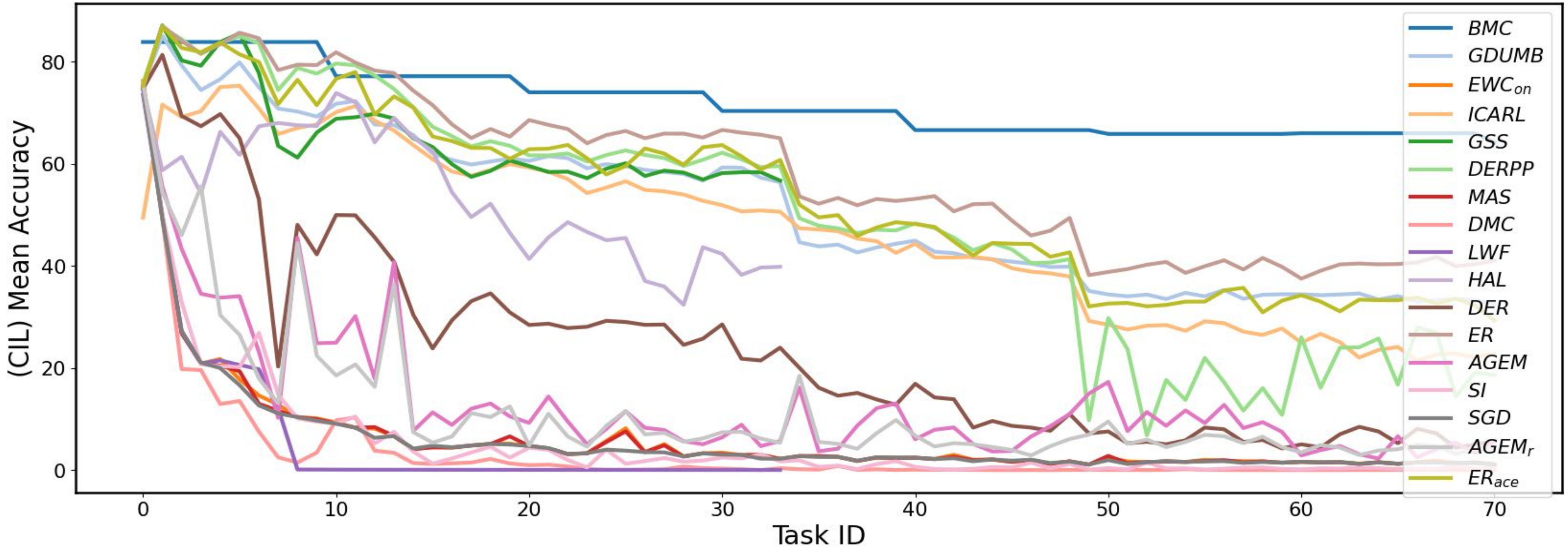
$$\mathcal{L}_{base} = \alpha\mathcal{L}_T(\theta'_{base}(x), y) + \boxed{\beta\mathcal{L}_{bmc}(\theta'_{base}, \mathcal{D})}$$

Distributed CL Training Framework



- Experts are trained individually on remote devices.
- Each remote device passes the Buffer data to central device once after expert training.
- The central device uses the Buffer and Memory data to update the Base Model.

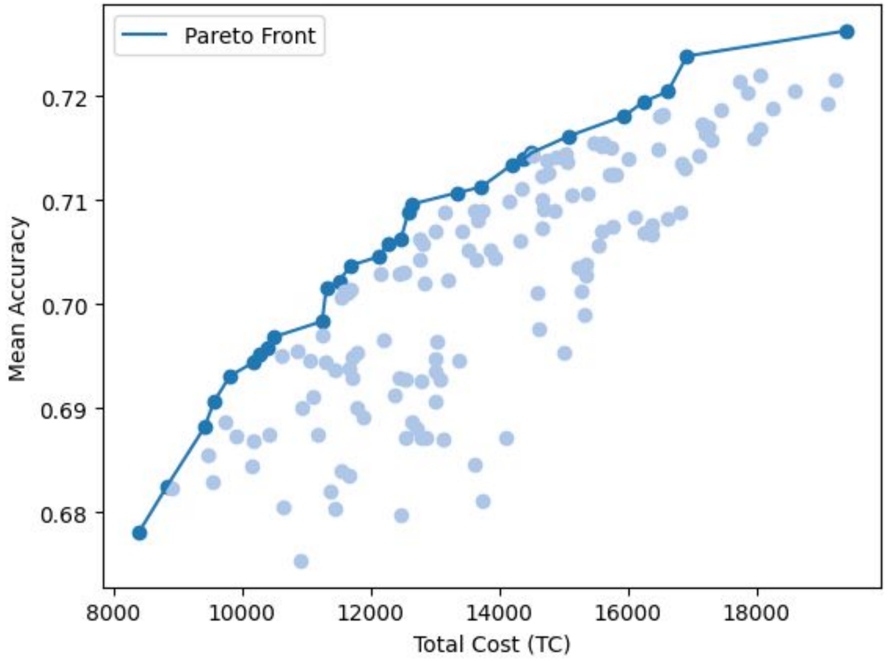
Experiments - Stream Benchmark

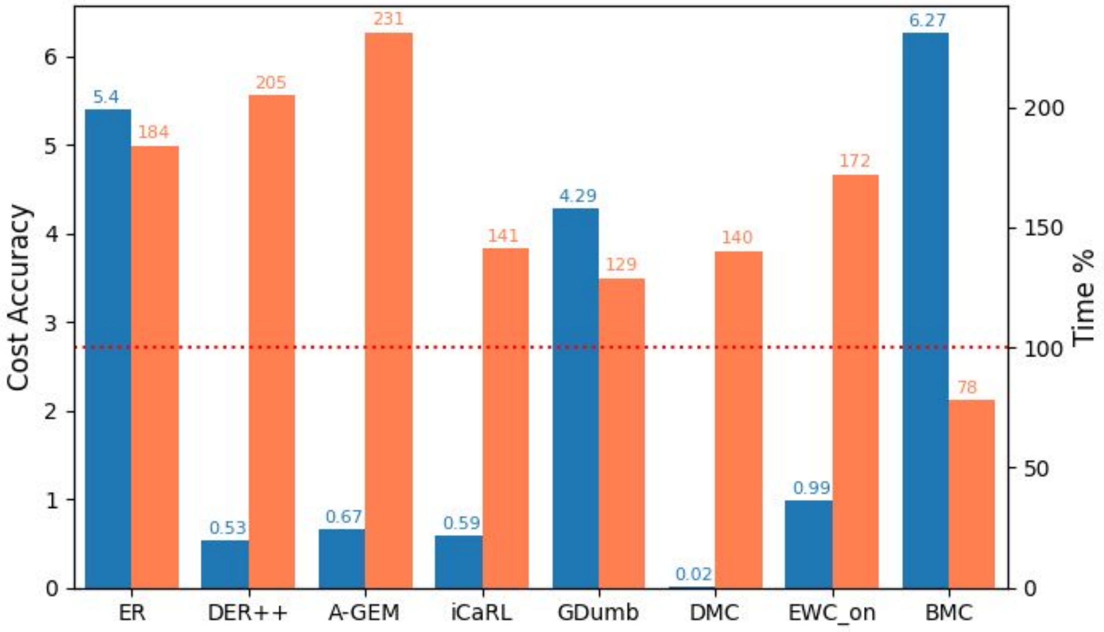


Stream dataset: 71 image classification datasets concatenated, with 6,770,722 training images, 743,977 validation images, and 2866 classes.

BMC achieved 70.4% final mean accuracy compared to the second best Experience Replay (ER) 41.4%, a 70% improvement.

USC





Experiments - Cost Analysis

Total Cost (TC)

= Communication Cost (buffer size)

+ Memory Cost (memory size)

The Pareto front of our method shows a trade-off between the Total Cost of a memory and a buffer with Mean Accuracy.

Cost Accuracy (blue)

The ratio between final mean accuracy and Total Cost, representing the performance gained per unit cost. BMC has the highest Cost Accuracy 6.27.

Relative Time Performance (orange)

The relative time of optimizing w.r.t. training sequentially without CL (SGD/fine-tuning). BMC has the highest time efficiency of 78% and is the only one faster than fine-tuning (100%, red dotted line).

Conclusions

- **BMC** is a combined approach of expert model regularization, rehearsal by experience replay, and parameter-isolation by training then consolidating disjoint experts.
- **BMC** allows distributed training where each expert reside on a different device and specialize in a given task.
- **BMC** is the only method that can maintains performance for our long sequence of 71 tasks, while being more efficient than sequential fine-tuning.
- A more sophisticated baseline such as DER++ [3] does not outperform Experience Replay in a more realistic dataset like **Stream**, calls for more research.

References

[1] Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

[2] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *CoRR*, abs/1909.08383, 2019.

[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, volume 33, pages 15920–15930. Curran Associates, Inc., 2020.