Supplementary Material for Heterogeneous Continual Learning

Divyam Madaan[1,2*], Hongxu Yin[1], Wonmin Byeon[1], Jan Kautz[1], Pavlo Molchanov[1]
[1]NVIDIA, [2]New York University
divyam.madaan@nyu.edu, {dannyy, wbyeon, jkautz, pmolchanov}@nvidia.com

Organization. In the supplementary material, we provide the implementation details and hyper-parameter configurations in Appendix A. Further, we show the results for class-IL continual learning and additional visualization of examples generated by QDI in Appendix B.

A. Experimental Details

We follow the base hyper-parameter setup from Buzzega et al. [1] for SCL and HCL experiments. We use an SGD optimizer for experiments with base learning rate as 0.03 for all the models. For each new task a total of 200 epochs are utilized to train the current model and we use the average accuracy metric on the validation set to store the best-model for evaluation for each task. Batch size is set as 32 and training is conducted on one NVIDIA V100 GPU of 16G for CIFAR-10 experiment and 32G for Split CIFAR-100 and Tiny-ImageNet datasets. For knowledge distillation methods, we use α and β equal to 1.0 for Split CIFAR-10, 3.0 for Split CIFAR-100 and Tiny-ImageNet and β equal to 0.3 with buffer. QDI previous class target labels are evenly sampled across previous tasks. Optimization hyper-parameters for QDI are provided next for three datasets, each shared across all networks during the training run:

- **Split CIFAR-10.** Optimization using Adam optimizer of learning rate 0.005,

  $$\alpha_{tv} = 0.001, \alpha_{\ell_2} = 0, \alpha_{\text{feature}} = 0.1$$

  for 0.5K iterations.$d(\cdot, \cdot)$ is MSE loss.

- **Split CIFAR-100.** Optimization using Adam with learning rate 0.03,

  $$\alpha_{tv} = 0.003, \alpha_{\ell_2} = 0.003, \alpha_{\text{feature}} = 0.2$$

  for 0.5K iterations. $d(\cdot, \cdot)$ is MSE loss.

- **Split Tiny-ImageNet.** Optimization based on Adam optimizer of learning rate 0.05,

  $$\alpha_{tv} = 0.001, \alpha_{\ell_2} = 0.05, \alpha_{\text{feature}} = 0.5$$

  for 0.5K iterations. $d(\cdot, \cdot)$ is MSE loss.

*Work done during an internship at NVIDIA.

For each dataset we found the QDI hyper-parameters based on a validation set obtainend by randomly sampling 10% of the training set. All results in main paper are on the test set, with 3 independent runs of the hyperset from random seeds for means and standard deviations.

B. Additional Results

Class-IL continual learning. Table B.1 shows the results for class-IL continual learning.

Visualization of synthesized examples. We provide extra visualization of QDI samples for the CIFAR-100 and Tiny-ImageNet datasets as in Fig. B.1. It can be observed that the proposed method scales across datasets with high fidelity in synthesized samples. More interestingly, the optimization step can find very close proxy of the current task semantics in older domains and dream out visual features of high perceptual realism.

References

[1] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[2] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[3] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[4] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

METHOD	$\mathcal{B}$	SPLIT CIFAR-10		SPLIT CIFAR-100		SPLIT TINY-IMAGENET	
		$\mathcal{A}_T$ (↑)	$\mathcal{F}_T$ (↓)	$\mathcal{A}_T$ (↑)	$\mathcal{F}_T$ (↓)	$\mathcal{A}_T$ (↑)	$\mathcal{F}_T$ (↓)
STANDARD CONTINUAL LEARNING							
FINETUNE	–	19.60 (± 0.03)	96.66 (± 0.12)	6.93 (± 1.13)	60.53 (± 1.11)	6.78 (± 0.14)	40.67 (± 0.92)
DI [6]	–	22.72 (± 1.02)	77.65 (± 0.30)	7.21 (± 0.75)	63.90 (± 2.87)	7.28 (± 2.05)	51.70 (± 0.63)
SI [7]*	–	19.48 (± 0.17)	95.78 (± 0.64)	–	–	6.58 (± 0.31)	–
LWF [3]*	–	19.61 (± 0.05)	96.69 (± 0.25)	–	–	8.46 (± 0.22)	–
KD (OURS)	–	**22.73** (± **0.75**)	87.00 (± 2.29)	16.77 (± 0.63)	80.65 (± 0.99)	20.86 (± 0.14)	42.38 (± 0.95)
KD W/ QDI (OURS)	–	19.75 (± 0.03)	**0.08** (± **0.02**)	**22.53** (± **1.15**)	**16.50** (± **2.72**)	**24.21** (± **0.52**)	**36.58** (± **1.56**)
ICARL [4]*	✓	49.02 (± 3.20)	28.72 (± 0.49)	–	–	7.53 (± 0.79)	–
A-GEM [2]	✓	21.98 (± 0.56)	92.18 (± 1.98)	5.04 (± 0.12)	91.93 (± 0.22)	7.49 (± 0.12)	72.04 (± 0.34)
ER [5]	✓	48.56 (± 1.40)	58.11 (± 1.61)	9.65(± 0.95)	85.20 (± 1.27)	10.70 (± 0.27)	83.99 (± 0.18)
DER [1]	✓	66.08 (± 1.18)	27.40 (± 2.16)	19.01 (± 0.74)	65.06 (± 0.15)	10.01 (± 1.52)	65.66 (± 3.60)
DER++ [1]	✓	67.23 (± 1.36)	26.13 (± 0.28)	21.38 (± 0.48)	56.17 (± 3.37)	8.23 (± 0.31)	68.51 (± 1.17)
KD W/ BUFFER (OURS)	✓	**70.34** (± 1.07)	**16.32** (± 0.18)	**25.19** (± **0.15**)	**40.53** (± **0.73**)	**29.21** (± **0.74**)	**16.48** (± **0.70**)
MULTITASK*	–	92.20 (± 0.15)	N/A	70.32 (± 0.48)	N/A	59.99 (± 0.19)	N/A
HETEROGENEOUS CONTINUAL LEARNING							
FINETUNE	–	21.45 (± 0.75)	91.24 (± 2.33)	5.27 (± 0.23)	72.99 (± 1.94)	7.90 (± 0.13)	57.23 (± 0.11)
DI [6]	–	20.67 (± 1.10)	70.97 (± 2.13)	6.20 (± 1.40)	73.70 (± 1.49)	6.39 (± 0.60)	51.14 (± 0.87)
KD (OURS)	–	30.21 (±0.11)	**27.54** (± **1.91**)	12.94 (± 1.13)	58.89 (± 1.64)	14.18 (± 0.35)	46.91 (± 0.59)
KD W/ QDI (OURS)	–	**33.89** (± **3.53**)	31.73 (± 3.51)	**15.86** (±**1.51**)	**32.73** (±**0.77**)	**15.38** (±**1.67**)	**37.27** (±**1.31**)
ER [2]	✓	38.77 (± 1.99)	69.61 (± 2.20)	7.43 (± 0.36)	82.85 (± 0.20)	7.59 (± 0.12)	61.77 (± 0.31)
A-GEM [2]	✓	19.67 (± 0.41)	89.26 (± 2.89)	4.62 (± 0.02)	88.78 (± 0.84)	6.93 (± 0.11)	63.93 (± 0.73)
DER [1]	✓	44.13 (± 0.98)	57.64 (± 4.76)	10.11 (± 0.65)	80.92 (± 1.35)	8.11 (± 0.26)	56.73 (± 0.81)
DER++ [1]	✓	48.82 (± 1.75)	50.11 (± 2.74)	10.97 (± 0.55)	73.62 (± 0.86)	8.88 (± 0.61)	55.54 (± 2.13)
KD W/ BUFFER (OURS)	✓	**65.40** (± **0.96**)	**10.13** (± **2.19**)	**21.00** (± **1.01**)	**38.49** (± **1.39**)	**18.77** (± **0.91**)	**8.76** (± **0.79**)

Table B.1. **Accuracy and forgetting** with class-IL on standard CL and HCL. The best results are highlighted in **bold**. $\mathcal{B}$ denotes replay-buffer, $\mathcal{A}_T, \mathcal{F}_T$ denote average accuracy and forgetting after the completion of training.* denotes the methods whose numbers were used from Buzzega et al. [1] and – indicates the unavailability of results. All other experiments are over three independent runs.

[5] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[6] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[7] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 2

Figure B.1. More QDI visualization for CIFAR-100 and Tiny-ImageNet datasets. Best viewed in color.