Supplementary Material:
Overcoming Catastrophic Forgetting with Unlabeled Data in the Wild

A. Illustration of Global Distillation

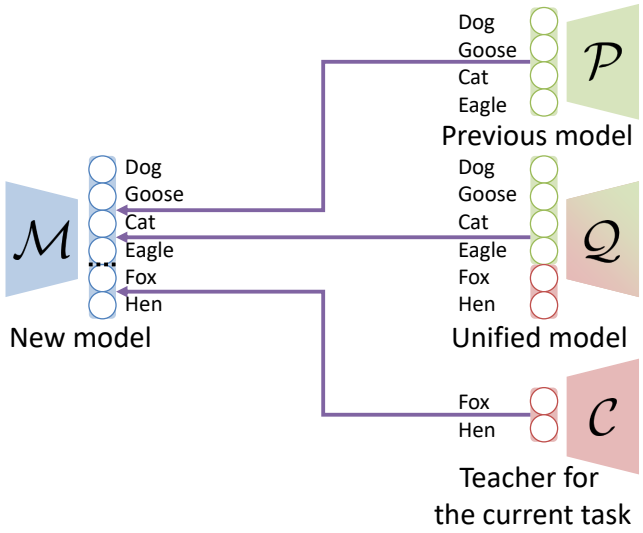


Figure A.1. An illustration of how a model $\mathcal{M}$ learns with global distillation (GD). For GD, three reference models are used: $\mathcal{P}$ is the previous model, $\mathcal{C}$ is the teacher for the current task, and $\mathcal{Q}$ is an ensemble of them.

B. Details on Experimental Setup

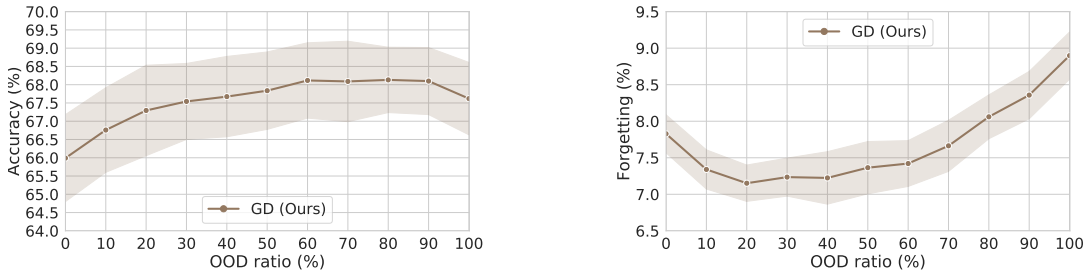
Hyperparameters. We use mini-batch training with a batch size of 128 over 200 epochs for each training to ensure convergence. The initial learning rate is 0.1 and decays by 0.1 after 120, 160, 180 epochs when there is no fine-tuning. When fine-tuning is applied, the model is first trained over 180 epochs where the learning rate decays after 120, 160, 170 epochs, and then fine-tuned over 20 epochs, where the learning rate starts at 0.01 and decays by 0.1 after 10, 15 epochs. We note that 20 epochs are enough for convergence even when fine-tuning the whole networks for some methods. We update the model parameters by stochastic gradient decent with a momentum 0.9 and an L2 weight decay of 0.0005. The size of the coreset is set to 2000. Due to the scalability issue, the size of the sampled external dataset is set to the size of the labeled dataset. The ratio of OOD data in sampling is determined by validation on a split of ImageNet, which is 0.7. For all experiments, the temperature for smoothing softmax probabilities is set to 2 for distillation from $\mathcal{P}$ and $\mathcal{C}$, and 1 for distillation from $\mathcal{Q}$. To be more specific about the way to scale probabilities, let $z = \{z_y | y \in \mathcal{T}\} = \mathcal{M}(x; \theta, \phi)$ be the set of outputs (or logits). Then, with a temperature γ, the probabilities are computed as follows:

$$p(y = k|x, \theta, \phi) = \frac{\exp(z_k/\gamma)}{\sum_{y' \in \mathcal{T}} \exp(z_{y'}/\gamma)}.$$

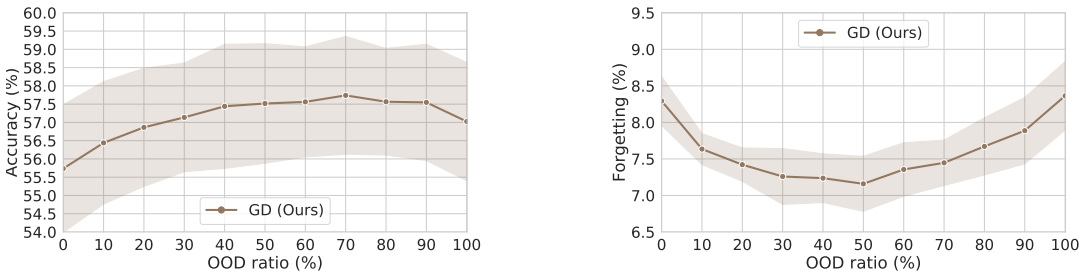
Scalability of methods. We note that all compared methods are scalable and they are compared in a fair condition. We do not compare generative replay methods with ours, because the coreset approach is known to outperform them in class-incremental learning in a scalable setting: in particular, it has been reported that continual learning for a generative model is a challenging problem on datasets of natural images like CIFAR-100 [3, 5].

C. More Experimental Results

C.1. More Ablation Studies



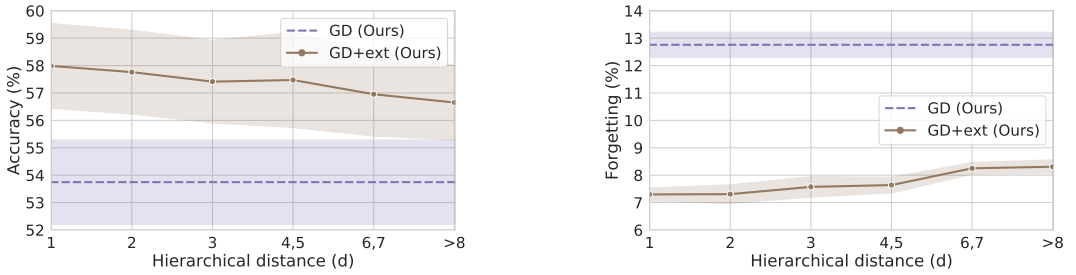
(a) CIFAR-100 ACC (b) CIFAR-100 FGT



(c) ImageNet ACC (d) ImageNet FGT

Figure C.1. Experimental results on CIFAR-100 and ImageNet when the task size is 10. We report ACC and FGT with respect to the OOD ratio averaged over ten trials for CIFAR-100 and nine trials for ImageNet.

Effect of the OOD ratio. We investigate the effect of the ratio between the sampled data likely to be in the previous tasks and OOD data. As shown in Figure C.1, the optimal OOD ratio varies over datasets, but it is higher than 0.5: specifically, the best ACC is achieved when the OOD ratio is 0.8 on CIFAR-100, and 0.7 on ImageNet. On the other hand, the optimal OOD ratio for FGT is different: specifically, the best FGT is achieved when the OOD ratio is 0.2 on CIFAR-100, and 0.5 on ImageNet.



(a) ImageNet ACC (b) ImageNet FGT

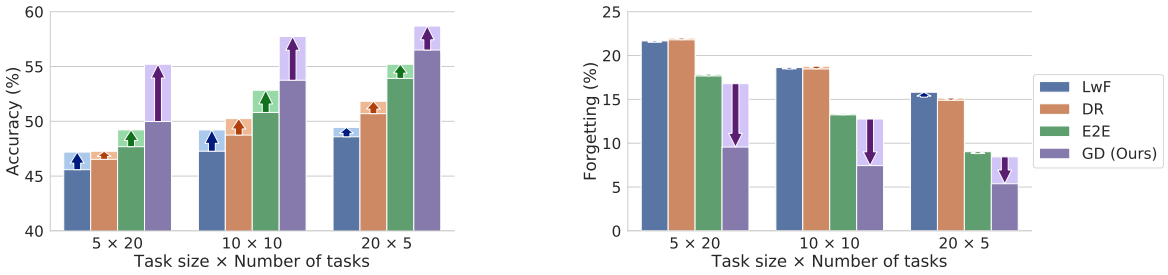
Figure C.2. Experimental results on ImageNet when the task size is 10. We report ACC and FGT with respect to the hierarchical distance between the training dataset and unlabeled data stream averaged over nine trials.

Effect of the correlation between the training data and unlabeled external data. So far, we do not assume any correlation between training data and external data. However, in this experiment, we control the correlation between them based on the hypernym-hyponym relationship between ImageNet class labels. Specifically, we first compute the hierarchical distance (the length of the shortest path between classes in hierarchy) between 1k classes in ImageNet ILSVRC 2012 training dataset and the other 21k classes in the entire ImageNet 2011 dataset. Note that the hierarchical distance can be thought as the semantic difference between classes. Then, we divide the 21k classes based on the hierarchical distance, such that each split has at least 1M images for simulating an unlabeled data stream. As shown in Figure C.2, the performance is proportional to the semantic similarity, which is inversely proportional the hierarchical distance. However, even in the worst case, unlabeled data are beneficial.

C.2. More Results

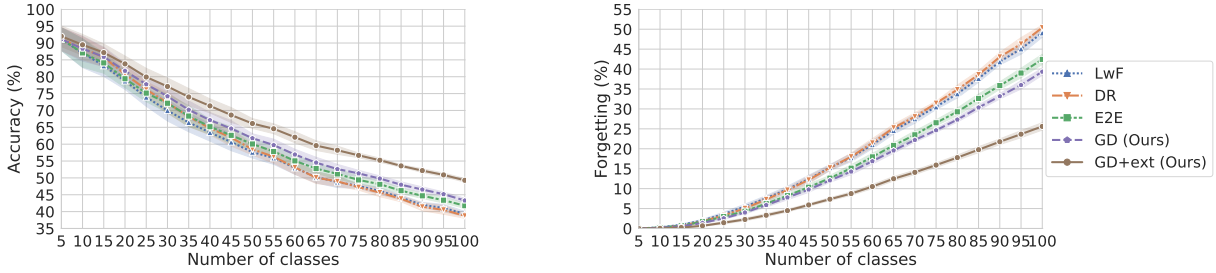
Table C.1. Comparison of methods on CIFAR-100 and ImageNet. We report the mean and standard deviation of ten trials for CIFAR-100 and nine trials for ImageNet with different random seeds in %. ↑ (↓) indicates that the higher (lower) number is the better.

Dataset	CIFAR-100						ImageNet					
Task size	5		10		20		5		10		20	
Metric	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
Oracle	78.6 ± 0.9	3.3 ± 0.2	77.6 ± 0.8	3.1 ± 0.2	75.7 ± 0.7	2.8 ± 0.2	68.0 ± 1.7	3.3 ± 0.2	66.9 ± 1.6	3.1 ± 0.3	65.1 ± 1.2	2.7 ± 0.2
Without an external dataset												
Baseline	57.4 ± 1.2	21.0 ± 0.5	56.8 ± 1.1	19.7 ± 0.4	56.0 ± 1.0	18.0 ± 0.3	44.2 ± 1.7	23.6 ± 0.4	44.1 ± 1.6	21.5 ± 0.5	44.7 ± 1.2	18.4 ± 0.5
LwF [4]	58.4 ± 1.3	19.3 ± 0.5	59.5 ± 1.2	16.9 ± 0.4	60.0 ± 1.0	14.5 ± 0.4	45.6 ± 1.9	21.5 ± 0.4	47.3 ± 1.5	18.5 ± 0.5	48.6 ± 1.2	15.3 ± 0.6
DR [2]	59.1 ± 1.4	19.6 ± 0.5	60.8 ± 1.2	17.1 ± 0.4	61.8 ± 0.9	14.3 ± 0.4	46.5 ± 1.6	22.0 ± 0.5	48.7 ± 1.6	18.8 ± 0.5	50.7 ± 1.2	15.1 ± 0.5
E2E [1]	60.2 ± 1.3	16.5 ± 0.5	62.6 ± 1.1	12.8 ± 0.4	65.1 ± 0.8	8.9 ± 0.2	47.7 ± 1.9	17.9 ± 0.4	50.8 ± 1.5	13.4 ± 0.4	53.9 ± 1.2	8.8 ± 0.3
GD (Ours)	**62.1 ± 1.2**	**15.4 ± 0.4**	**65.0 ± 1.1**	**12.1 ± 0.3**	**67.1 ± 0.9**	**8.5 ± 0.3**	**50.0 ± 1.7**	**16.8 ± 0.4**	**53.7 ± 1.5**	**12.8 ± 0.5**	**56.5 ± 1.2**	**8.4 ± 0.4**
With an external dataset												
LwF [4]	59.7 ± 1.2	19.4 ± 0.5	61.2 ± 1.1	17.0 ± 0.4	60.8 ± 0.9	14.8 ± 0.4	47.2 ± 1.7	21.7 ± 0.5	49.2 ± 1.3	18.6 ± 0.4	49.4 ± 0.8	15.8 ± 0.4
DR [2]	59.8 ± 1.0	19.5 ± 0.5	62.0 ± 0.9	16.8 ± 0.4	63.0 ± 1.0	13.9 ± 0.4	47.3 ± 1.7	21.8 ± 0.6	50.2 ± 1.5	18.5 ± 0.5	51.8 ± 0.9	14.9 ± 0.5
E2E [1]	61.5 ± 1.2	16.4 ± 0.5	64.3 ± 1.0	12.7 ± 0.4	66.1 ± 0.9	9.2 ± 0.4	49.2 ± 1.7	17.7 ± 0.6	52.8 ± 1.4	13.2 ± 0.2	55.2 ± 0.9	9.0 ± 0.4
GD (Ours)	**66.3 ± 1.2**	**9.8 ± 0.3**	**68.1 ± 1.1**	**7.7 ± 0.3**	**68.9 ± 1.0**	**5.5 ± 0.4**	**55.2 ± 1.8**	**9.6 ± 0.4**	**57.7 ± 1.6**	**7.4 ± 0.3**	**58.7 ± 1.2**	**5.4 ± 0.3**



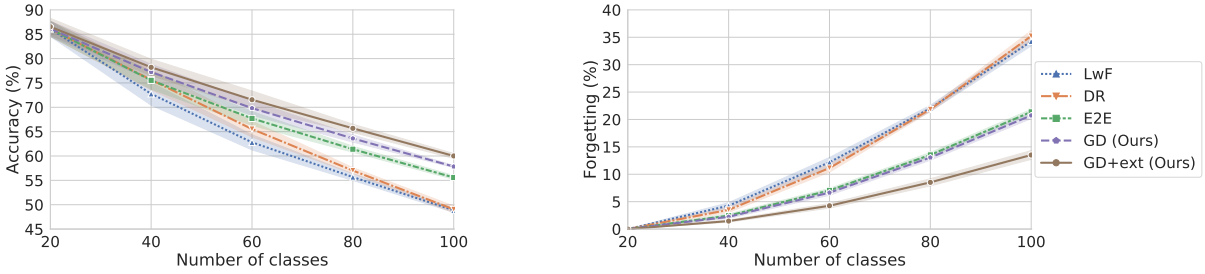
(a) ACC improvement by learning with external data (b) FGT improvement by learning with external data

Figure C.3. Experimental results on ImageNet. Arrows show the performance gain in ACC and FGT by learning with unlabeled data, respectively. We report the average performance of nine trials.



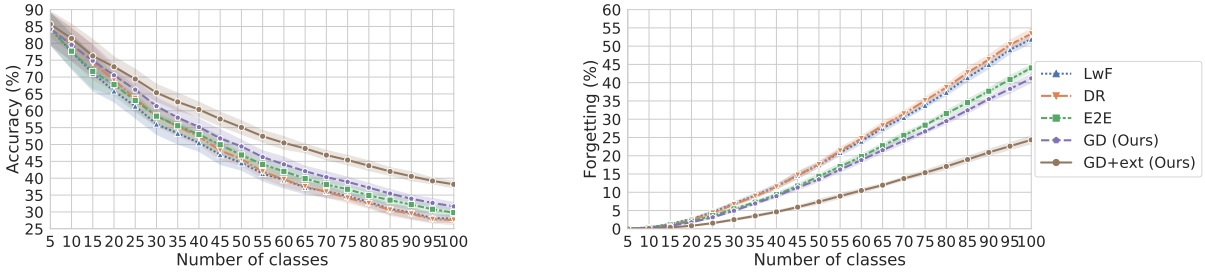
(a) ACC with respect to the number of trained classes (b) FGT with respect to the number of trained classes

Figure C.4. Experimental results on CIFAR-100 when the task size is 5. We report ACC and FGT with respect to the number of trained classes averaged over ten trials.



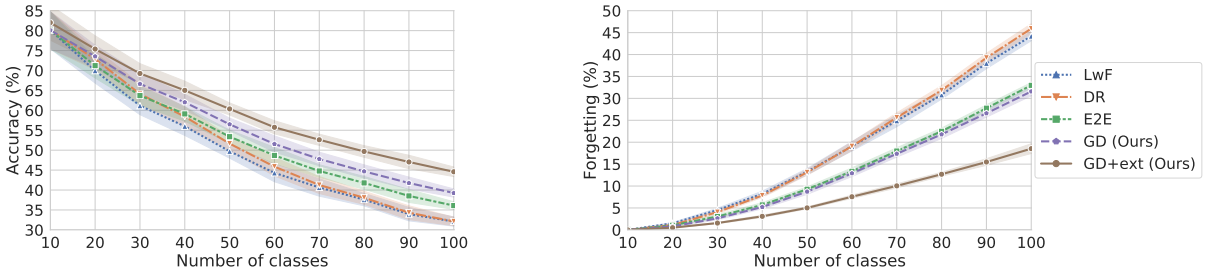
(a) ACC with respect to the number of trained classes (b) FGT with respect to the number of trained classes

Figure C.5. Experimental results on CIFAR-100 when the task size is 20. We report ACC and FGT with respect to the number of trained classes averaged over ten trials.



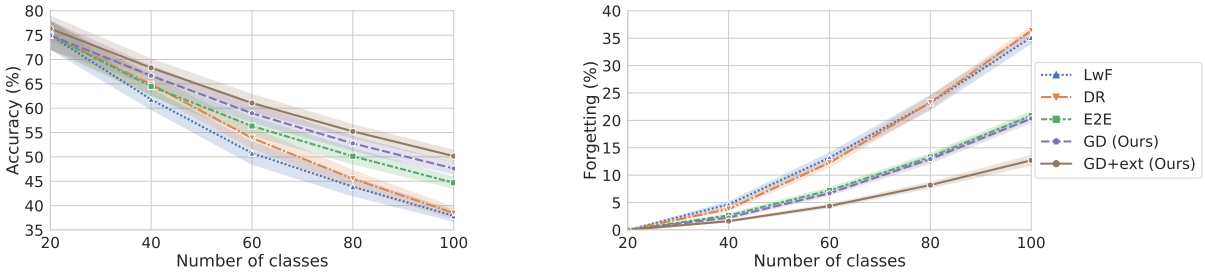
(a) ACC with respect to the number of trained classes (b) FGT with respect to the number of trained classes

Figure C.6. Experimental results on ImageNet when the task size is 5. We report ACC and FGT with respect to the number of trained classes averaged over nine trials.



(a) ACC with respect to the number of trained classes (b) FGT with respect to the number of trained classes

Figure C.7. Experimental results on ImageNet when the task size is 10. We report ACC and FGT with respect to the number of trained classes averaged over nine trials.



(a) ACC with respect to the number of trained classes (b) FGT with respect to the number of trained classes

Figure C.8. Experimental results on ImageNet when the task size is 20. We report ACC and FGT with respect to the number of trained classes averaged over nine trials.

References

[1] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *ECCV*, 2018.

[2] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Lifelong learning via progressive distillation and retrospection. In *ECCV*, 2018.

[3] T. Lesort, H. Caselles-Dupré, M. Garcia-Ortiz, A. Stoian, and D. Filliat. Generative models from the perspective of continual learning. *arXiv preprint arXiv:1812.09111*, 2018.

[4] Z. Li and D. Hoiem. Learning without forgetting. In *ECCV*, 2016.

[5] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, Z. Zhang, and Y. Fu. Incremental classifier learning with generative adversarial networks. *arXiv preprint arXiv:1802.00853*, 2018.