Class-incremental Learning via Deep Model Consolidation
Supplemental Material Document

Junting Zhang[1] Jie Zhang[2] Shalini Ghosh[3] Dawei Li[3] Serafettin Tasci[3] Larry Heck[3]
Heming Zhang[1] C.-C. Jay Kuo[1]
[1]University of Southern California [2]Arizona State University [3]Samsung Research America

Appendix Overview

In this supplemental document, we provide additional detailed experimental results and analyses of the proposed method, Deep Model Consolidation (DMC), for class-incremental learning.

A. Detailed experimental results of DMC for object detection

In the experiments of DMC for incremental learning of object detectors, we incrementally learn $19+1$ classes using RetinaNet [7]. In the main paper, we presented the results of adding "tvmonitor" class as the new class. Here, we show the results of addition of one class experiment with each of the VOC categories being the new class in Table 5, where *Old Model* denotes the 19-class detector trained on the old 19 classes, *New Model* denotes the 1-class detector trained on the new class and *DMC* denotes the final consolidated model that is capable of detecting all the 20 classes. Per-class average precisions on the entire test set of PASCAL VOC 2007 [4] are reported.

B. Effect of the amount of auxiliary data for object detection

We studied the effect of the amount of auxiliary data for DMC for image classification task. To see how the amount of auxiliary data affects the final performance in the incremental learning of object detection, we performed additional experiments on PASCAL VOC 2007 with the $10+10$ classes setting. We randomly sampled $1/2$, $1/4$ and $1/8$ of the full auxiliary data from Microsoft COCO dataset [10] for consolidation. As shown in Table 1, with just $1/8$ of full data, *i.e.*, 12.3k images, DMC can still outperform the state-of-the-art, which demonstrates its robustness and efficiency in the detection task as well.

C. Implementation and training details

We implement DMC with PyTorch [8] library.

Table 1. Varying the amount of auxiliary data in the consolidation stage. VOC 2007 test mAP (%) are shown, where classes 1-10 are the old classes, and classes 11-20 are the new ones.

Model	Old Classes	New Classes	All Classes
All auxiliary data	70.53	66.16	68.35
$1/2$ of auxiliary data	69.79	66.06	67.93
$1/4$ of auxiliary data	70.2	64.67	67.44
$1/8$ of auxiliary data	66.77	62.71	64.74

Training details for the image classification experiments. Following iCaRL [9], we use a 32-layers ResNet [5] for all experiments and the model weights are randomly initialized. When training the individual specialist models, we use SGD optimizer with momentum for 200 epochs. In the consolidation stage, we train the network for 50 epochs. The learning rate schedule for all the experiments is same, *i.e.*, it starts with 0.1 and reduced by 0.1 at 7/10 and 9/10 of all epochs. For all experiments, we train the network using mini-batches of size 128 and a weight decay factor of 1×10^{-4} and momentum of 0.9. We apply the simple data augmentation for training: 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded image or its horizontal flip.

Training details for the object detection experiments. We resize each image so that the smaller side has 640 pixels, keeping the aspect ratio unchanged. We train each model for 100 epochs and use Adam [6] optimizer with learning rate 1×10^{-3} on two NVIDIA Tesla M40 GPUs simultaneously, with batch size of 12. Random horizontal flipping is used for data augmentation. Standard non-maximum suppression (NMS) with threshold 0.5 is applied for post-processing at test time to remove the duplicate predictions. For each image, we select 64 anchor boxes for DMC training. Empirically we found selecting more anchor boxes (128, 256 etc.) did not provide further performance gain. The λ is set to 1.0 for all experiments.

Hyperparameters used for the baseline methods. We report results of *EWC++* [3], *SI* [11], *MAS* [1] and *RWalk* [3] on iCIFAR-100 benchmark in the main paper. Table 2 summarizes the hyperparamter $\hat{\lambda}$ that controls the strength of regularization used in each experiment, and they are picked based on a held-out validation set.

Methods	$g=5$	$g=10$	$g=20$	$g=50$
EWC++ [3]	10	10	1	0.1
SI [11]	0.01	0.05	0.01	0.01
MAS [1]	0.1	0.1	0.001	0.0001
RWalk [3]	5	1	1	0.1

Table 2. $\hat{\lambda}$ used in when incrementally learning g classes at a time on iCIFAR-100 benchmark.

D. Preliminary experiments of adding exemplars

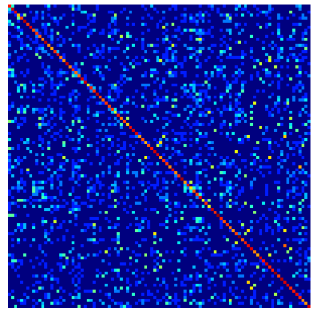
While DMC is realistic in applied scenarios due to its scalablity and immunity to copyright and privacy issues, we additionally tested our method in the scenario where we are allowed to store some exemplars from the old data with a fixed budget when learning the new classes. Suppose we are incrementally learning a group of g classes at a time, With the same total memory budget $K = 2000$ as in iCaRL [9], we fill the exemplar set by randomly sampling $\lfloor \frac{K}{g} \rfloor$ training images from each class when we learn the first group of classes; then every time we learn g more classes with training data $\mathcal{D}_{new} = [X^{gi}, \cdots, X^{g(i+1)-1}]$ in the i-th incremental learning session, we augment the exemplar set by $\lfloor \frac{K}{gi} \rfloor$ randomly sampled training images of the new classes, and we fine-tune the consolidated model using these exemplars for 15 epochs with a small learning rate of 1×10^{-3}. After fine-tuning, we reduce the size of the exemplar set by keeping $\lfloor \frac{K}{g(i+1)} \rfloor$ exemplars for each class. We refer to this variant of our method as DMC+. We validate the effectiveness of DMC+ on the iCIFAR-100 benchmark, and Table 3 summarizes the results as the average of the classification accuracies over all steps of the incremental training (as in [2], the accuracy of the first group is not considered in this average). We can get comparable performance to iCaRL in all settings. Note that we also tried the herding algorithm to select the exemplars as in iCaRL, but we did not observe any notable improvement.

The confusion matrices comparison between DMC+ and iCaRL [9] is shown in Fig. 1, and we find: 1) fine-tuning with exemplars can indeed further reduce the intrinsic bias in the training; 2) our DMC+ is on a par with iCaRL, even though we use naive random sampling rather than the more expensive herding [9] approach to select exemplars.

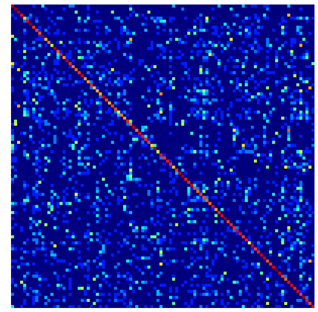
These preliminary results demonstrate that DMC may also hold promise for exemplar-based incremental learning, and we would like to further study the potential improvement of DMC+, *e.g.* in terms of exemplar selection scheme and rehearsal strategies.

Table 3. Average incremental accuracies when adding the exemplars of old classes. iCaRL [9] with the same memory budget is compared. Results of incremental learning with $g = 5, 10, 20, 50$ classes at a time on iCIFAR-100 benchmark are reported.

g	5	10	20	50
iCaRL	$\mathbf{57.8 \pm 2.6}$	$\mathbf{60.5 \pm 1.6}$	62.0 ± 1.2	61.8 ± 0.4
DMC+	56.78 ± 0.86	59.1 ± 1.4	$\mathbf{63.2 \pm 1.3}$	$\mathbf{63.1 \pm 0.54}$



(a) DMC+



(b) iCaRL

Figure 1. Confusion matrices of exemplar-based methods on iCIFAR-100 when incrementally learning 10 classes in a group. The element in the i-th row and j-th column indicates the percentage of samples with ground-truth label i that are classified into class j. Fig. 1(b) is from [9]. (Best viewed in color.)

E. Preliminary experiments of consolidating models with common classes

The original DMC assumes the two models to be consolidated are trained with distinct sets of classes, but it can be easily extended to the case where we have two models that are trained with partially overlapped set of classes. We first normalize the logits produced by the two models as Eq. 4 in the main paper. We then set the double distillation regression target as the follows: for the common classes, we take the mean of normalized logits from the two model; for each of the other classes, we take the normalized logit from the corresponding specialist model that was trained with this class.

Below we present a preliminary experiment on CIFAR-100 dataset in this setting, where we have separately trained two 55-class classifiers for *Class 1-55* and *Class 46-100*, respectively, where 10 classes (*Class 46-55*) are in common. The results are shown in Table 4. For the common classes, DMC can be considered as an ensemble learning method, where at least the accuracy of the weaker model is maintained; for learning the rest of classes, it does not exhibit catastrophic forgetting or intransigence. This shows that DMC is promisingly extensible to the special case of incremental learning with partially overlapped categories.

Table 4. Consolidation of two models with 10 common classes (class 46-55).

Model	Class 1-45	Class 46-55	Class 56-100	Class 1-100
Model 1	73.73	80.5	-	-
Model 2	-	71.6	66.47	-
Consolidated	60.76	71.7	58.09	60.65

F. Enlarged plots

We provide enlarged plots of accuracy curves for iCIFAR-100 ($g = 5, 10, 20, 50$) in Fig. 2 for better visibility.

References

[1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 1, 2

[2] F. M. Castro, M. J. Marın-Jiménez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[3] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[6] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 1

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1

[8] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 1

[9] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[10] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 1

[11] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3987–3995. JMLR. org, 2017. 1, 2

Table 5. VOC 2007 test per-class average precision (%) when incrementally learning 19 + 1 classes.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Old Model	-	78.8	77.4	56.5	60.1	76.4	85.0	80.0	50.0	78.0	69.9	78.3	79.2	74.3	77.3	39.5	66.4	65.7	76.9	74.4	-
New Model	15.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	16.3	75.9	75.8	52.9	59.5	74.2	84.2	79.1	49.3	73.0	59.9	70.0	75.4	64.8	79.9	40.2	64.1	58.8	69.9	74.3	64.9
Old Model	69.6	-	76.3	60.1	59.8	76.7	85.4	79.6	54.6	75.9	63.7	78.6	79.5	71.5	77.7	44.9	68.0	57.6	77.3	75.5	-
New Model	-	70.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	75.8	62.4	75.5	59.6	59.0	76.0	85.6	79.5	53.8	77.0	62.3	78.6	77.6	67.5	80.7	43.5	70.6	57.6	77.3	76.3	69.8
Old Model	68.9	78.8	-	55.9	61.4	70.7	79.9	79.8	50.9	73.6	65.0	77.7	79.3	76.0	77.0	43.1	66.2	66.8	77.3	75.5	-
New Model	-	-	35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	69.0	77.9	43.8	54.7	60.1	75.5	84.1	77.5	51.0	71.4	65.4	69.4	69.7	73.5	76.5	40.8	59.9	66.9	77.0	76.2	67.0
Old Model	76.9	78.3	77.1	-	57.9	76.2	85.2	79.8	48.7	76.5	65.6	82.9	76.9	75.1	77.7	40.6	67.7	67.6	76.9	69.5	-
New Model	-	-	-	18.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	76.6	77.2	75.8	23.4	58.2	77.2	84.4	80.0	48.7	78.5	63.3	82.8	70.3	76.1	80.7	40.8	66.7	64.9	75.5	68.6	68.5
Old Model	70.5	77.9	77.5	53.5	-	76.1	85.6	78.8	51.0	76.2	62.5	77.2	79.1	73.2	77.6	42.5	68.6	68.1	76.6	74.5	-
New Model	-	-	-	-	47.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	74.7	76.2	76.4	51.0	37.6	76.9	85.4	79.4	53.0	76.7	64.2	77.9	77.9	73.6	80.4	43.0	68.2	68.4	76.5	75.3	69.6
Old Model	70.8	77.8	75.2	57.2	60.0	-	84.7	79.6	48.3	75.3	68.4	78.8	78.6	75.6	77.3	41.8	69.0	68.0	75.0	73.9	-
New Model	-	-	-	-	-	46	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	68.7	79.7	73.9	55.6	61.3	53.5	84.9	79.3	49.4	75.8	66.8	78.9	75.6	75.4	80.6	41.6	67.4	66.7	70.0	73.8	68.9
Old Model	77.5	78.8	74.5	58.1	60.3	74.5	-	80.7	49.0	76.0	64.4	77.3	78.7	66.8	77.1	39.0	67.9	67.0	77.1	75.3	-
New Model	-	-	-	-	-	-	76.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	70.3	76.3	74.0	51.3	60.2	68.2	77.5	80.0	47.1	76.5	61.0	77.4	77.3	59.5	79.9	41.5	66.5	65.0	77.2	74.8	68.1
Old Model	76.5	79.4	78.1	54.7	60.8	77.2	85.4	-	49.6	74.9	65.1	78.5	78.5	74.3	77.8	44.2	67.3	65.1	76.0	74.5	-
New Model	-	-	-	-	-	-	-	60.5	-	-	-	-	-	-	-	-	-	-	-	-	-
DMC	75.7	81.0	76.6	51.4	61.9	76.7	84.5	69.8	51.5	74.6	63.6	76.9	69.4	74.6	81.2	43.3	67.0	67.1	77.1	74.2	69.9
Old Model	78.7	79.6	76.9	57.3	62.2	77.4	80.0	79.5	-	75.9	66.6	77.6	79.6	76.5	77.3	43.4	67.3	66.7	77.8	69.3	-
New Model	-	-	-	-	-	-	-	-	41.9	-	-	-	-	-	-	-	-	-	-	-	-
DMC	75.8	76.2	76.9	56.7	62.9	76.9	85.4	78.9	38.1	75.1	64.1	78.6	76.0	74.4	80.4	43.0	66.8	62.6	77.8	74.0	70.0
Old Model	70.8	77.8	76.0	58.1	60.7	78.1	85.0	80.1	47.2	-	64.4	77.4	75.3	74.9	80.3	41.7	66.8	64.9	77.1	72.1	-
New Model	-	-	-	-	-	-	-	-	-	30.3	-	-	-	-	-	-	-	-	-	-	-
DMC	69.9	75.6	68.1	56.4	60.7	77.2	85.5	79.4	46.4	37.0	65.2	70.0	68.0	74.6	80.4	41.7	59.6	62.8	76.5	72.9	66.4
Old Model	75.5	80.1	77.1	57.8	61.4	76.6	85.5	80.6	51.1	79.0	-	78.6	80.2	75.4	77.1	44.7	68.4	66.7	77.4	74.6	-
New Model	-	-	-	-	-	-	-	-	-	-	43.6	-	-	-	-	-	-	-	-	-	-
DMC	75.0	80.8	75.3	54.1	62.6	76.8	85.3	80.6	50.2	77.4	53.9	83.8	77.6	73.4	81.0	45.9	66.8	65.7	75.4	74.8	70.8
Old Model	76.6	77.8	77.2	57.4	60.6	76.3	84.8	80.9	49.9	77.4	64.5	-	77.4	69.3	77.5	43.2	73.7	67.4	76.7	74.7	-
New Model	-	-	-	-	-	-	-	-	-	-	-	40.3	-	-	-	-	-	-	-	-	-
DMC	75.2	76.6	74.5	57.1	62.0	74.9	85.4	70.0	51.0	57.8	63.6	52.5	59.2	73.5	79.9	43.1	65.4	66.9	75.1	74.5	66.9
Old Model	77.3	78.2	77.7	59.4	60.5	77.5	85.2	85.9	48.8	76.6	70.7	76.5	-	74.1	77.3	42.1	67.8	68.0	78.7	72.4	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	52.4	-	-	-	-	-	-	-	-
DMC	77.4	76.2	72.1	54.9	63.0	77.5	84.7	79.1	48.0	73.3	68.0	61.5	40.5	71.9	79.5	40.4	65.9	63.0	77.4	73.0	67.4
Old Model	76.5	77.3	75.7	56.8	60.8	70.5	85.4	79.7	48.6	74.2	62.7	79.3	77.3	-	76.9	43.9	68.4	63.3	77.2	76.0	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	59.2	-	-	-	-	-	-	-
DMC	68.9	74.2	75.8	55.0	60.1	77.1	84.2	86.0	50.7	75.2	61.3	78.8	70.0	68.2	79.6	46.1	68.1	61.4	75.5	76.1	69.6
Old Model	77.1	79.7	76.9	59.1	62.3	77.3	85.7	80.2	52.0	77.6	65.0	78.5	80.4	78.2	-	44.1	67.5	71.9	78.0	74.1	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76.4	-	-	-	-	-	-
DMC	75.8	78.8	75.0	59.7	62.1	77.1	85.5	80.1	51.1	77.0	63.3	77.9	78.1	76.8	78.0	44.7	66.0	69.2	78.1	75.1	71.5
Old Model	75.5	77.1	75.5	58.9	62.1	77.8	85.8	87.7	44.4	76.6	64.7	78.3	78.7	75.5	77.4	-	68.3	67.7	76.6	73.4	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.8	-	-	-	-	-
DMC	75.5	74.3	74.4	56.1	61.6	77.9	86.7	87.2	48.0	77.1	64.6	78.0	77.4	73.7	80.3	31.0	67.3	65.8	75.4	73.8	70.3
Old Model	78.0	77.8	75.3	57.5	61.0	69.6	86.5	79.8	48.5	67.9	62.8	76.8	79.6	74.8	77.5	43.5	-	68.0	76.8	74.8	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26	-	-	-	-
DMC	76.3	76.9	73.6	54.3	62.0	73.8	86.1	79.7	48.5	67.0	63.9	75.6	76.3	75.2	80.8	44.4	20.2	68.2	76.4	74.4	67.7
Old Model	77.6	78.2	76.3	55.0	59.3	70.7	85.8	80.4	50.5	75.4	67.2	83.5	78.7	69.0	77.6	44.4	67.7	-	70.1	75.1	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33.1	-	-	-
DMC	77.4	82.5	68.5	58.3	61.8	75.2	85.6	78.7	47.1	74.9	63.5	75.6	69.8	73.5	79.4	42.6	65.8	26.1	69.8	74.1	67.5
Old Model	70.4	79.5	77.1	57.6	60.2	73.6	84.6	80.2	51.0	75.5	65.4	78.7	78.0	75.3	77.7	42.8	69.3	63.6	-	73.9	-
New Model	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	34.9	-	-
DMC	73.9	77.4	76.7	56.0	60.8	62.1	83.7	79.7	49.6	76.3	65.7	79.3	73.8	73.2	80.6	40.3	73.3	65.9	37.7	74.5	68.0

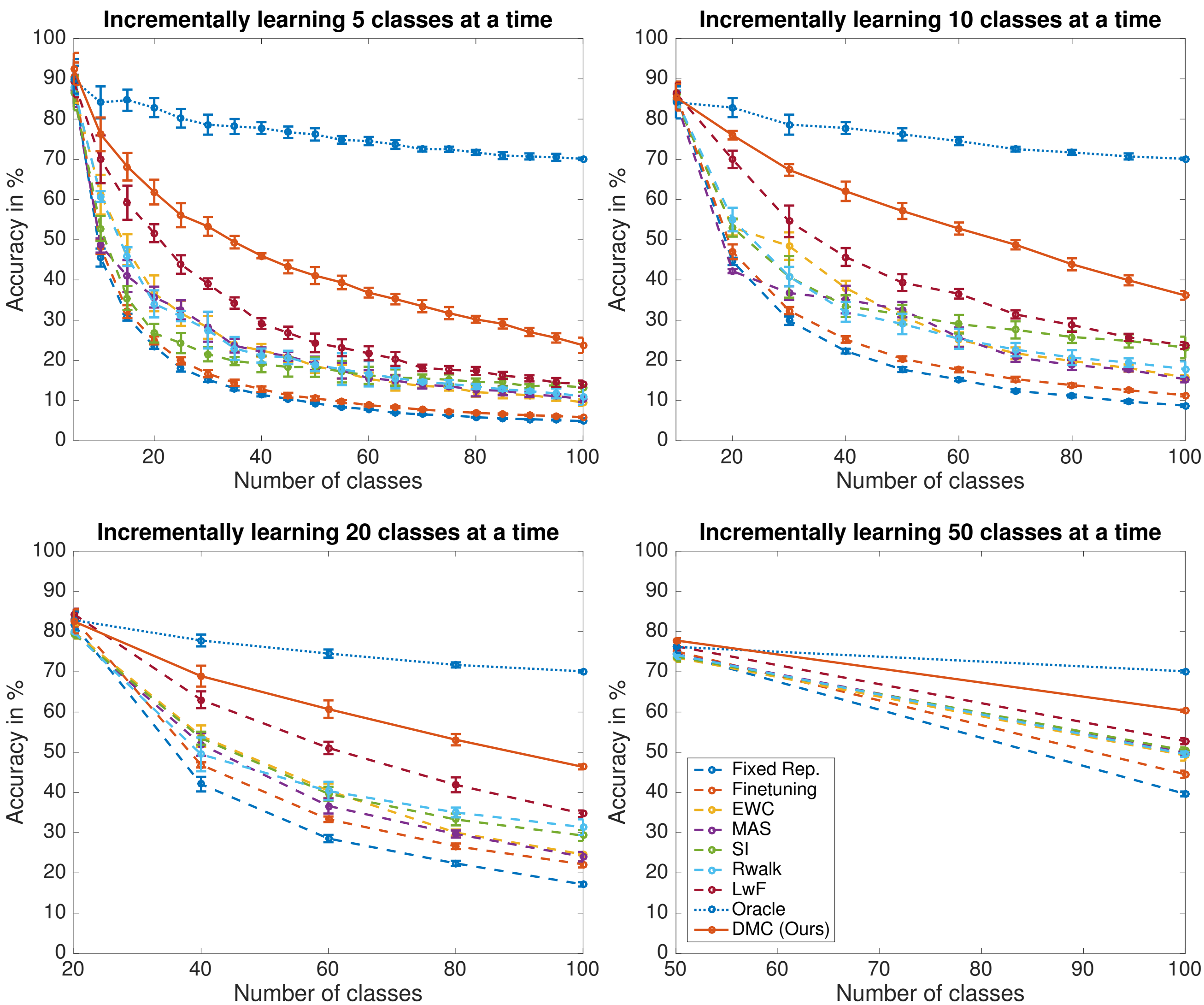


Figure 2. Incremental learning with group of $g = 5, 10, 20, 50$ classes at a time on iCIFAR-100 benchmark.