M^2KD: Incremental Learning via Multi-model and Multi-level Knowledge Distillation

Peng Zhou[1]
pengzhou@umd.edu
Long Mai[2]
malong@adobe.com
Jianming Zhang[2]
jianmzha@adobe.com
Ning Xu[2]
nxu@adobe.com
Zuxuan Wu[1]
zxwu@cs.umd.edu
Larry S. Davis[1]
lsd@umiacs.umd.edu

[1] University of Maryland
College Park, MD. USA.
[2] Adobe Research
345 Park Avenue
San Jose, CA. USA.

A Appendix

A.1 Dataset Details

iILSVRC-small [5]: A small subset of 100 classes out of the 1000-class ImageNet. The evaluation metric is the 5 iteration average accuracy of the 100 classes. The number of classes in every incremental step is determined by the number of steps. We train from scratch and the performance is evaluated on the validation set of ImageNet.
Cifar-100 [7]: All 100 classes are evaluated. The average result out of 5 random selections of classes for each incremental step is reported. We train from scratch and the evaluation results are based on the test set.

A.2 Ablation Studies

We investigate the effectiveness of each component of our method in this section. In particular, we compare our full model with the following baselines.
LWF-MC aux: Add auxiliary distillation to LWF-MC.
LWF-MC MMD: Change the original loss to our multi-model distillation. No auxiliary distillation is applied.
Ours skip1: Instead of using all previous models, we study the case when skipping some snapshots. Starting from the last previous model, we skip the first model in multi-model distillation. The skipped model is replaced by the next model for multi-model distillation.
Ours skip2: Skip the first two models instead of one compared to **Ours skip1**.

© 2020. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

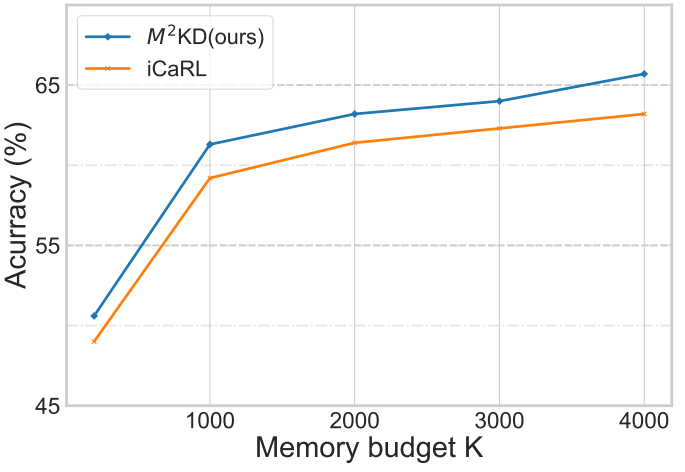
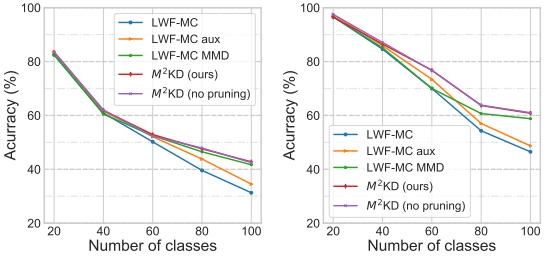


Figure 1: Analysis on performance and memory compared to iCaRL on Cifar-100 (10-class batch).



(a) Top-1 Cifar-100 (b) Top-5 iILSVRC-small

Figure 2: Ablation Studies for our approach (20-class batch).

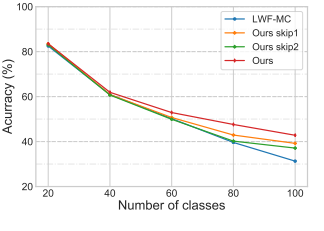


Figure 3: Comparison between different number of models used in multi-model distillation on Cifar-100 20-class batch.

Figure 2 shows the comparison for each of the component in our approach. **LWF-MC aux** improves our baseline model **LWF-MC** on all the datasets after adding auxiliary distillation, indicating that the intermediate level information also contributes to preserving previous knowledge. With only multi-model distillation (**LWF-MC MMD**), the performance gradually improves for both datasets as more incremental steps are involved, which demonstrates that directly distilling knowledge from the corresponding model helps to reduce the lost in sequential distillation. Note that our multi-model distillation reduces to the standard distillation used in [4] if only one or two incremental steps are added. By incorporating the auxiliary distillation, however, our method still shows improved performance. Lastly, our model achieves nearly the same performance as our upper bound which saves all previous snapshots, showing the effectiveness of our pruning based approach.

Figure 3 compares how multi-model distillation is affected by the number of models. **LWF-MC** can be regarded as a special case which skips 3 models in the last round. The trend from **LWF-MC** to **Ours** shows that the performance improves as the number of model preserved increases, confirming the value of multi-model distillation.

	20-batch	10-batch	5-batch
EEIL [1]	65.1	64.5	63.7
BiC [5]	66.2	65.2	64.0
Ours	66.2	64.8	64.3

Table 1: Additional comparison for exemplar-based setting on Cifar-100. Each entry is the average top-1 accuracy under different incremental batches.

	Cifar-100	iLSVRC-small
iCaRL	80	40
EEIL [1]	80 + 30	40 + 30
LwM [4]	-	40 + 40
Ours	80 + 15	40 + 15

Table 2: Training time comparison. Each entry is the number of training epochs in Cifar-100 and iLSVRC-small. '-' denotes that the result is not available in the literature.

A.3 Further Memory Analysis

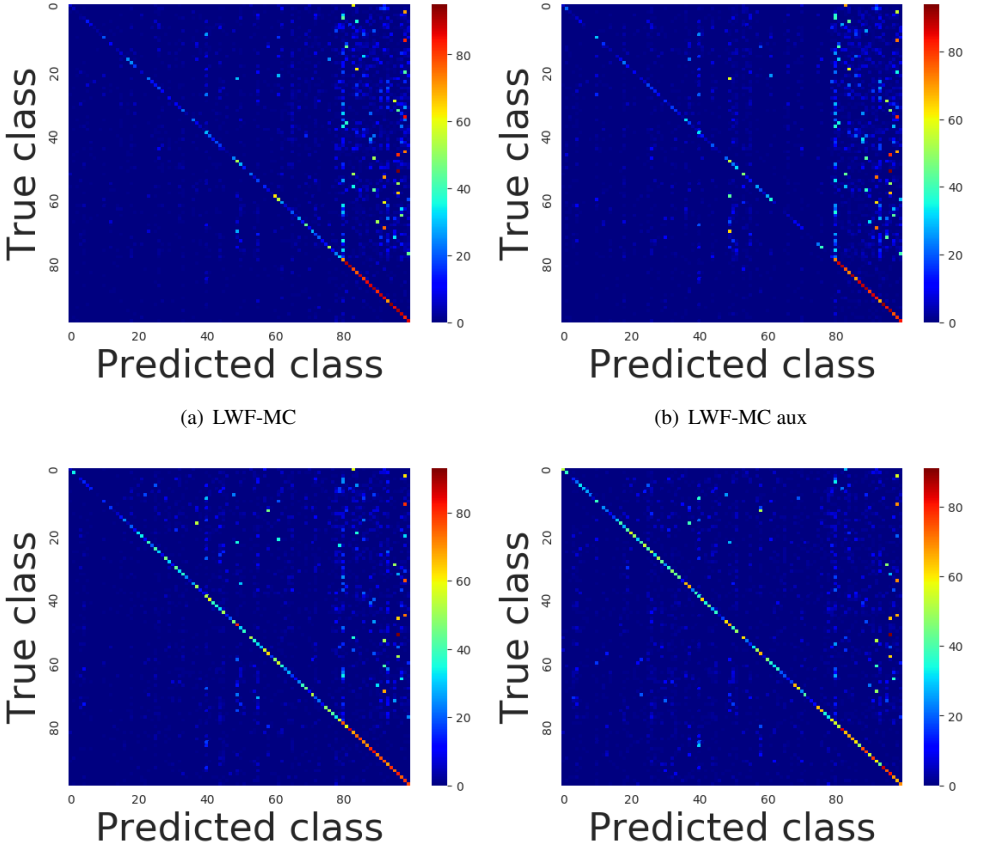
We provide further memory analysis in Figure 1. We compare our approach with iCaRL on Cifar-100 given the same memory constraint. For fair comparison, we reduce the exemplar set as a penalty of the additional memory we use for network parameters to match with the memory size used for iCaRL. The performance is evaluated by averaging the top-1 accuracy across all the incremental steps. When memory budget equals to 200 images, we do not use any exemplar set but still perform better than iCaRL. The reason for this is that the sequential distillation pipeline tends to lose information even when exemplars from old classes are available. Moreover, increasing memory budget makes the performance gap between our approach and iCaRL larger, showing our strength to memorize what has been learned. Please see **supplementary material** for more analysis on training overhead.

A.4 Additional Exemplar-based Comparison

We compare the performance in exemplar-based setting on Cifar-100 with SOTA approaches [1, 5] and list the results in Table 1. It shows that our approach achieves comparable results. Note that replacing the sequential distillation in these approaches with our method could boost their performance further.

A.5 Training Overhead

Our approach only requires small overhead in training and no additional cost at inference compared with LWF and iCaRL. Actually, additional training is a common strategy to balance data in incremental learning. Table 2 lists the training epoch comparison. It shows that our overhead is less than the balance fine-tuning overhead in [1] and the sample generation in [4], and we need only 18.8% and 37.5% additional training time than iCaRL. Even though we use all the previous footprints to get pseudo labels, similar in iCaRL, one strategy is to calculate the previous outputs for all the training images before each incremental step and use them as labels during training. The overhead of label calculation on a GPU 1080Ti is 2.3 seconds per incremental step on Cifar-100 and 6.0 seconds on iLSVRC-small (10-class batch), which is negligible. For testing, no additional computation is required since only one model is used.



(a) LWF-MC (b) LWF-MC aux

(c) Ours skip2 (d) M^2KD (ours)

Figure 4: Confusion matrix comparison for Cifar-100 in exemplar-free setting. (20-class batch) **(a)** LWF-MC. **(b)** LWF-MC with auxiliary distillation. **(c)** Ours skip2. **(d)** M^2KD (ours).

A.6 Confusion Matrix Visualization

We show the resulting confusion matrices in Figure 4. It can be observed from the confusion matrix that **LWF-MC** has a strong bias to the data from the newly added classes while the performance on the old classes degrades dramatically. With the knowledge from the intermediate level, the confusion of previous data gets reduced. More clearly, with the favor of multi-model distillation, the knowledge from all the previous data preserves better and cause less confusion. Also, if we skip some previous models in the distillation and use other models to guide the network, the skipped logits become less confident than directly using the corresponding model for distillation. In short, the comparison from confusion matrices confirms the advantage on preserving previous knowledge via multi-model distillation.

References

[1] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018.

[2] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[3] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2018.

[4] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *CVPR*, 2019.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[6] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019.