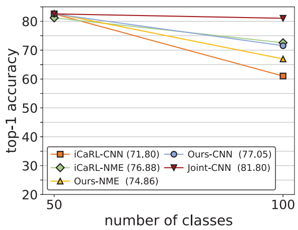
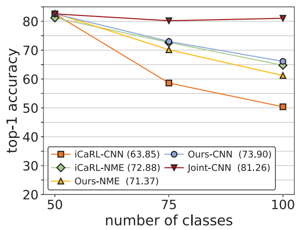
Learning a Unified Classifier Incrementally via Rebalancing

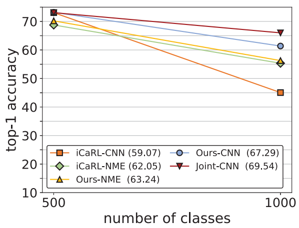
Saihui Hou[1*], Xinyu Pan[2*], Chen Change Loy[3], Zilei Wang[1], Dahua Lin[2]
[1] University of Science and Technology of China, [2] The Chinese University of Hong Kong,
[3] Nanyang Technological University
saihui@mail.ustc.edu.cn, px118@ie.cuhk.edu.hk, ccloy@ntu.edu.sg,
zlwang@ustc.edu.cn, dhlin@ie.cuhk.edu.hk



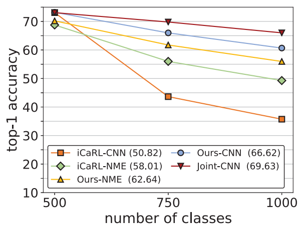
(a) ImageNet-Subset (1 phase)



(b) ImageNet-Subset (2 phases)

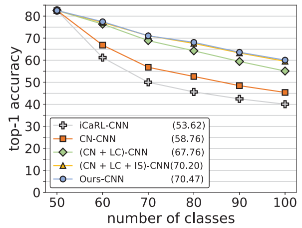


(c) ImageNet-Full (1 phase)

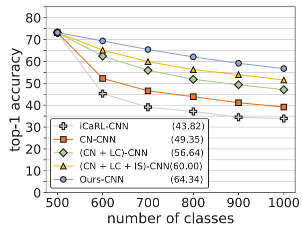


(d) ImageNet-Full (2 phases)

Figure 1. More evaluation on ImageNet. Reported on ImageNet-Subset (100 classes) and ImageNet-Full (1000 classes).



(a) ImageNet-Subset (5 phases)



(b) ImageNet-Full (5 phases)

Figure 2. The effect of each component on ImageNet. Reported on ImageNet-Subset (100 classes) and ImageNet-Full (1000 classes)

1. Appendix

1.1. More evaluation on ImageNet

Figure 1 shows the performance comparison on ImageNet in one phase and two phases. The results indicate that our approach also performs better (*Ours-CNN vs. iCaRL-NME*) under these settings.

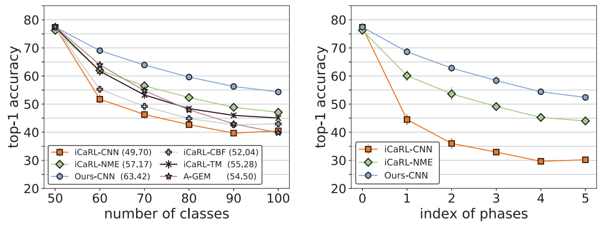
1.2. The effect of each component on ImageNet

Figure 2 illustrates the effect of each component on ImageNet. It is worth noting that, *CBF* makes a little contribution on ImageNet-Subset as it does on CIFAR100. However, it leads to a considerable improvement on ImageNet-Full, where the performance of our method without *CBF* is already significantly better than those by iCaRL. One possible reason is that, the new data coming at each phase on ImageNet-Full is ten times of that on ImageNet-Subset, resulting in the more severe imbalance between old and new classes on ImageNet-Full.

1.3. More experimental comparison

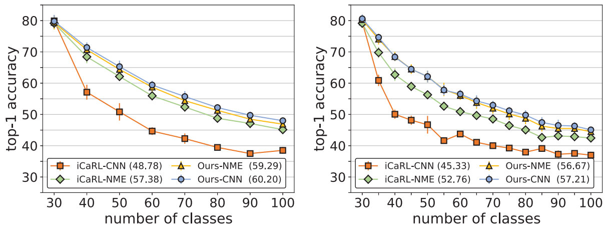
Figure 3(a) provides the results compared to more baselines. Specifically, Castro *et al.* [2] build on iCaRL and add a *class balance finetune* (denoted as *iCaRL-CBF*) on the reserved samples for all classes. Its performance without the sophisticated data augmentation is inferior to *iCaRL-NME* which agrees with the ablation study in [2]. Javed *et al.* [4] apply *threshold moving* [1] (denoted as *iCaRL-TM*) to the CNN prediction of iCaRL, while the performance is still a little inferior to *iCaRL-NME*. The more recent work *A-GEM* [3] is an improved version of *GEM* [5], both of which are proposed for the *multi-task* setting. We follow the idea of projected gradient and re-implement *A-GEM* under the *multi-class* setting. From the results in Figure 3(a), we can observe that, the proposed method outperforms those baselines as well as *iCaRL-NME* by a large margin.

Besides, Figure 3(b) shows the accuracy curves on the first batch of classes by different methods under the incremental setting of 5 phases on CIFAR100. The accuracy curve by *Ours-CNN* goes down more slowly, indicating that the previous knowledge is more effectively preserved in the proposed method.



(a) CIFAR100 (5 phases) (b) CIFAR100 (5 phases)

Figure 3. (a) The comparison with more baselines. (b) The accuracy curves on the first batch of classes.



(a) CIFAR100 (7 phases) (b) CIFAR100 (14 phases)

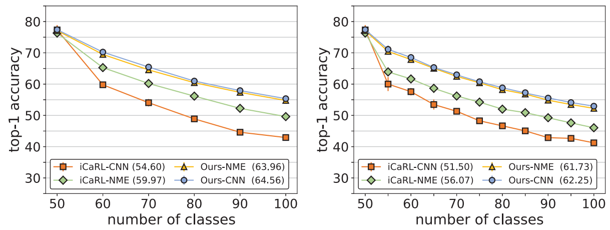
Figure 4. The results with different number of initial classes (30 classes on CIFAR100). The remaining 70 classes come in 7 and 14 phases (*i.e*. 10 and 5 classes at a time).

1.4. The results with different number of initial classes

In the paper our experiments start from a model trained on half of the classes on each dataset, and the settings are consistent with those in real-world applications where incremental learning usually starts from a model trained on a pre-collected dataset [6]. Here we provide some results with different number of initial classes, *e.g*. 30 classes on CIFAR100, where our method also performs better than the baselines as shown in Figure 4. It's worth mentioning that, the distillation loss $L_{\text{dis}}^{\text{G}}$ is computed on the features, which prefers a more representative feature extractor in the original model.

1.5. The results with fixed memory

Figure 5 illustrates the performance comparison with fixed memory, *i.e*. a memory with fixed capacity (*e.g*. $R_{\text{total}} = 2000$ for CIFAR100) is utilized to reserve the samples for old classes. It can be seen that, both iCaRL and our approach perform better with this strategy to reserve old samples (compared to $R_{\text{per}} = 20$ for CIFAR100), while our approach is also superior to iCaRL under different settings.



(a) CIFAR100 (5 phases) (b) CIFAR100 (10 phases)

Figure 5. The performance on CIFAR100 with fixed memory ($R_{\text{total}} = 2000$) to reserve old samples.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

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

[3] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *ICLR*, 2019.

[4] Khurram Javed and Faisal Shafait. Revisiting distillation and incremental classifier learning. *arXiv preprint arXiv:1807.02802*, 2018.

[5] David Lopez-Paz et al. Gradient episodic memory for continual learning. In *NIPS*, 2017.

[6] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *arXiv preprint arXiv:1802.07569*, 2018.