Reducing Class-wise Confusion for Incremental Learning with Disentangled Manifolds

Supplementary Material

1. Additional experiments

We provide an additional comparison on the fine-grained dataset CUB-200 in Table 1. CUB-200 [3] has 200 classes, with 11,788 images in total. We utilize ResNet18 (without pretraining) as the backbone, and set the memory buffer to 20 exemplars per class for all methods. We evaluate the proposed method on B0 Inc20 and B100 Inc20. In a highly similar scenario, our method surpasses other methods. It outperforms PsHD (NeurIPS' 2024) [1] up to 4.88% on average accuracy, which benefits from the proposed architecture and the separation loss that jointly learn class-specific subspaces.

Table 1. Comparison on a fine-grained dataset CUB-200.

Methods	B0 Inc20			B100 Inc20		
	#P	Last	Avg	#P	Last	Avg
BEEF	111.70	49.07	56.15	67.02	63.73	66.11
DSGD	111.70	58.10	59.75	67.02	67.35	68.80
PsHD	111.70	57.29	59.77	67.02	68.07	68.97
CREATE	17.72	**58.78**	**60.81**	17.72	**70.82**	**73.85**

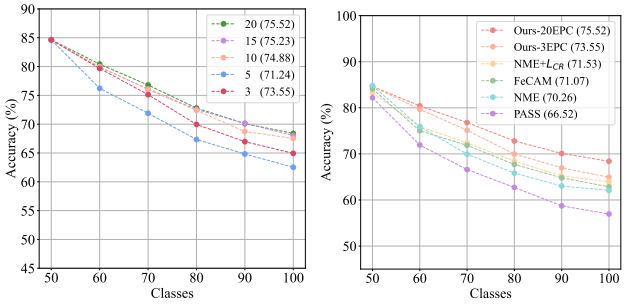
2. Additional ablation study

2.1. Impact of exemplar size

We conduct additional ablation experiments to evaluate the performance by varying the size of the exemplar set. In the CIFAR100 Base50 Inc10 setting, we record the performance with 20, 10, 5, and 3 exemplars stored for each class, as shown in Fig. 1(a). When the number of exemplars per class (EPC) is reduced from 20 to 10, the final accuracy decreases from 68.4% to 67.53%, indicating stable performance. When the EPC is dramatically reduced to only 3 exemplars per class, the average accuracy drops from 75.52% to 73.55%, resulting in a tolerable decrease of 1.97%.

2.2. Comparisons with exemplar-free methods

Our proposed method is efficient and can achieve competitive performance compared with exemplar-free methods when adapted to lower storage or cost requirements. FeCAM [2] is the SOTA of the exemplar-free methods, which proposes a feature covariance-aware metric based on prototypes for CIL. It stores a covariance matrix for each old class (106 MB in total), therefore, its memory cost is significantly higher than that of our proposed method CREATE when storing 3 exemplars per class (0.87 MB in total). PASS [4] freezes the backbone after the initial learning phase and stores one prototype for each class in the feature space. It requires 0.19 MB of memory but demands additional computation to implement self-supervised learning, augments rotation-based transformations for new samples, and adds Gaussian noise for prototypes. The experimental results are shown in Fig. 1(b). In the common exemplar-based setting, Ours-20EPC (20 exemplars per class) achieves a gain of 4.45% and 9.0% compared with FeCAM and PASS. Additionally, when converted to the corresponding memory overhead, Ours-3EPC (3 exemplars per class) shows a 2.48% improvement over FeCAM with a 99.18% reduction in storage cost and a 7.03% improvement over PASS with lower costs to learn new classes.



(a) exemplars per class (b) baselines

Figure 1. Additional ablation study on different exemplar sizes and other baselines. The values in the parentheses in the legend represent the average accuracy. "EPC" is an abbreviation for "exemplars per class".

We also analyze the effectiveness of the L_{CR} loss in the feature space without the reconstruction module, referred to as $NME + L_{CR}$. Experimental results indicate that using the L_{CR} loss independently in feature space can also improve the performance of CIL by 1.27%.

2.3. Analysis of stability and plasticity

To evaluate the performance improvements stemming from enhanced knowledge acquisition or reduced forgetting of the framework, we compare DER, BEEF, and our proposed method. The experiments are conducted using the CIFAR100 Base50 Inc50 set-up, where the number of old classes is the same as that of new classes. We evaluate the accuracy of old classes, new classes, and the overall accuracy. The experimental results are presented in Fig. 2. Our proposed method shows a 1.54% gap in learning new classes compared to previous methods, but it achieves approximately a 5% improvement in retaining old classes,

significantly reducing the forgetting of old classes. Our method achieves a better balance between stability and plasticity.

3. Pseudo code

In Algorithm 1, we present the pseudo code for our proposed method CREATE.

Algorithm 1 CREATE

Require: Dataset set $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_T\}$, seen class number in each phase $C = \{C_1, C_2, ..., C_T\}$, memory buffer $\mathcal{M}_t$, feature extractor ϕ_t, classifier θ_t, and auto-encoder module AE_i.

for task t ∈ [1, 2, ..., T] **do**
  if t==1 **then**
    Training set $\hat{\mathcal{D}}_t \leftarrow \mathcal{D}_1$
    Create AE_i for new class $i, i = \{0, ..., C_1\}$
    Calculate cross-entropy loss L_{CE} ▷ Eq. (3)
    Calculate confusion-reduce loss L_{CR} ▷ Eq. (8)
    Train ϕ_t and θ_t by loss $L = L_{CE} + L_{CR}$
  else
    Training set $\hat{\mathcal{D}}_t \leftarrow \mathcal{D}_t \bigcup \mathcal{M}_t$
    Freeze ϕ_{t-1}, θ_{t-1}, unfreeze ϕ_t,
    Create AE_i for new class $i, i = \{C_{t-1}, ..., C_t\}$
    Calculate cross-entropy loss L_{CE} ▷ Eq. (3)
    Calculate knowledge distillation loss L_{KD} according to the logits of $(\phi_{t-1}, \theta_{t-1})$ ▷ Eq. (4)
    Calculate confusion-reduce loss L_{CR} ▷ Eq. (8)
    Train ϕ_t and θ_t by loss $L = L_{CE} + L_{KD} + \lambda L_{CR}$ ▷ Eq. (9)
    Freeze ϕ_t
    Training set $\mathcal{D}_t^{'} \leftarrow$ sample a class-balanced subset from M_{t-1} and D_t
    Fine-tune θ_t by Eq. (9)
  end if
  Old feature extractor $\phi_{t-1} \leftarrow \phi_t$
  Old classifier $\theta_{t-1} \leftarrow \theta_t$
  $\mathcal{M}_{t-1} \leftarrow \mathcal{M}_t$
end for

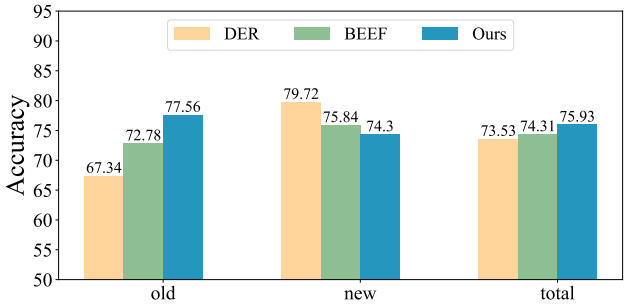


Figure 2. The accuracy of old, new, and total classes on CIFAR100 Base50 Inc50. Our method achieves performance gains by mitigating forgetting more effectively than other methods.

References

[1] Yan Fan, Yu Wang, Pengfei Zhu, Dongyue Chen, and Qinghua Hu. Persistence homology distillation for semi-supervised continual learning. In *Advances of Neural Information Processing Systems*, 2024. 1

[2] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[3] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[4] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 1