Supplementary Materials for "Beyond Prompt Learning: Continual Adapter for Efficient Rehearsal-Free Continual Learning"

Xinyuan Gao[1#], Songlin Dong[2#], Yuhang He[2⋆], Qiang Wang[1], and Yihong Gong[1,2]

[1] School of Software Engineering, Xi'an Jiaotong University
[2] College of Artificial Intelligence, Xi'an Jiaotong University
{gxy010317,dsl97273141,qwang}@stu.xjtu.edu.cn, heyuhang@xjtu.edu.cn
ygong@mail.xjtu.edu.cn

A Appendix

A.1 Discuss with LoRA and Naive Adapter

LoRA [5] and adapter [1, 4] aim to tuning the model to adapt the downstream task with a few of parameters. LoRA [5] is composed of two different low-rank matrices and is attached in parallel with the QKV attention. Adaptformer [1] proposes that attaching the adapter in parallel with MLP leads to better performance. However, they lack a viable mechanism (similar to key-query matching in prompt learning) to address the issue of the absence of task identifiers during the inference stage, cannot effectively resolve the RFCL problem. Compared with classical adapter tuning and LoRA, C-ADA expands the trainable parameters in CAL for every new task, which is parameter-extensible to solve continual tasks. Joint training of old and new weights in CAL allows the knowledge of the new task to evolve from the old task and reduces forgetting old knowledge. Besides, the attached position of C-ADA is diverse, C-ADA can be attached to different positions such as QKV attention, MLP, projection layer, etc.

A.2 Compare with Adapter in Continual Learning

There are some previous works [2, 10] introduce the adapter tuning into the continual learning field. [2] maintains a memory buffer (rehearsal-based method) and leverages a distillation mechanism to merge the adapters from different tasks to keep the stored parameters unchanged. [10] adapts a single adapter to tuning the model in the first task. For subsequent tasks, adapter and backbone are frozen and only the classifier is trained. Differ from the previous works, C-ADA is a parameter-extensible architecture to solve the RFCL task. Assigning a sub-adapter to each task and using orthogonality to reduce conflicts leads to better performance. We provide a more comprehensive result in table 1.

⋆ Yuhang He is the corresponding author; # Xinyuan Gao and Songlin Dong are co-first authors

Table 1: Results (%) on ImageNet-R. The results are all obtained by CODA [7], APG [8] and ADAM [10] directly. A_N is the final (last) accuracy over N tasks.

Methods	P=5		P=10		P=20	
	A_N (↑)	*Param* (↓)	A_N (↑)	*Param* (↓)	A_N (↑)	*Param* (↓)
L2P	70.83	0.7/100.7	69.29	0.7/100.7	65.89	0.7/100.7
Deep L2P	73.93	9.6/109.6	71.66	9.6/109.6	68.42	9.6/109.6
DualPrompt	73.05	0.5/100.5	71.32	0.8/100.8	67.87	1.3/101.3
CODA-P-S	75.19	0.7/100.7	73.93	0.7/100.7	70.53	0.7/100.7
CODA-P	76.51	4.6/104.6	75.45	4.6/104.6	72.37	4.6/104.6
APG	72.36	–	73.27	–	71.22	–
ADAM	74.23	0.8/100.9	72.87	0.8/100.9	70.47	0.8/100.9
C-ADA(d=1)	**76.23**	0.3/100.3	**75.06**	0.3/100.3	**71.92**	0.3/100.3
C-ADA(d=3)	**77.93**	0.7/100.7	**76.66**	0.7/100.7	**73.47**	0.7/100.7
C-ADA(d=5)	**78.53**	1.1/101.1	**76.91**	1.1/101.1	**73.72**	1.1/101.1

A.3 Further Comparison with CODA

In reality, CODA [7] has a lightweight model CODA-P-S (0.7%) and a larger model CODA-P (4.6%). For a more comprehensive comparison with CODA, we illustrate the relationship between trainable parameters and accuracy (A_N) in Figure 1. It is noteworthy that despite CODA's attempt to enhance performance through parameter expansion, it still falls significantly short of our C-ADA. Remarkably, C-ADA can utilize only **1/10** of the parameters of CODA-P to achieve superior performance by improving the model's plasticity. Moreover, the precision advantage of C-ADA over CODA-P expands with the increase of parameter number.

This finding is crucial as it demonstrates a significant advantage of C-ADA in terms of parameter efficiency. This is particularly important for resource-constrained environments and applications, which require minimizing the number of parameters while maintaining high performance. Overall, our results strongly suggest that C-ADA outperforms CODA in both parameter efficiency and performance, providing valuable insights for future research in computer vision.

A.4 Compared with the Generative Replay Methods

In the related work, we introduced a type of generative replay method [3, 6, 9]. They employ a GAN network to generate images of previous classes to mitigate the forgetting of the model to solve the RFCL problem. We report the result in table 2. Despite these methods being rehearsal-free, they typically underperform when compared to the prompt learning method (Our C-ADA), experiencing a decline in performance of approximately **30%**.

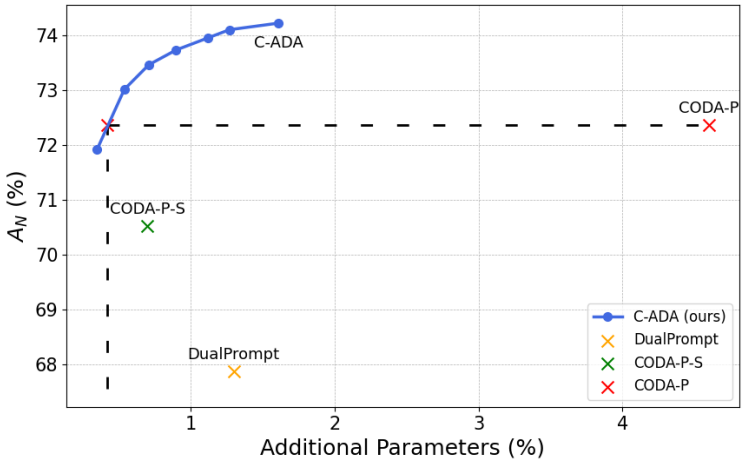

Fig. 1: The Comparison on Accuracy and Additional Parameter.

Table 2: The experiment is conducted on CIFAR-100 10 tasks. All results are obtained from IRP. [9].

Methods	**ABD** [6]	**R-DFCIL** [3]	**IRP** [9]	**C-ADA**
Acc(%)	59.0	61.7	68.3	87.18

A.5 Extra Overhead

The primary objective of C-ADA is to solve RFCL tasks with a lightweight and efficient framework. We present the additional overhead of ViT-B16 and C-ADA in table 3. Remarkably, by adding very few parameters (**0.7%**) and training complexity (**0.05%**), C-ADA exhibits outstanding competence in tackling the CL problem, which validates the lightweight and efficiency of C-ADA in the RFCL setting.

Table 3: The extra training and inference overhead on ImageNet-R 10 tasks.

Methods	**FLOPs Ratio**	**Iteration Time Ratio**	**#Param Ratio**
ViT-B16	100.00%	100.00%	100.00%
C-ADA	100.05%	103.19%	100.70%

A.6 Sensitive Study of Hyperparameters

To search for the best choice of the hyperparameter, we conduct the experiments on ImageNet-R 10 tasks with the different hyperparameter $\lambda \in (0, 01, 1)$ and

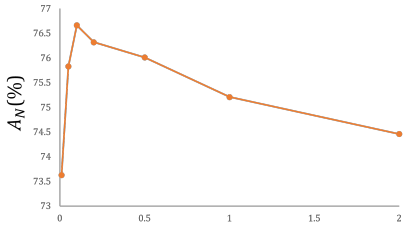

Fig. 2: The result of different hyperparameter $\lambda \in (0, 01, 1)$ on ImageNet-R 10 tasks.

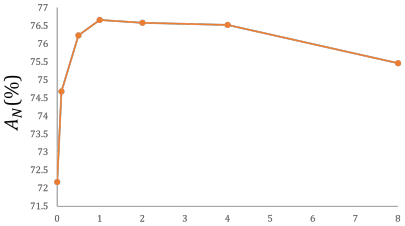

Fig. 3: The result of different hyperparameter $\delta \in (0, 8)$ on ImageNet-R 10 tasks.

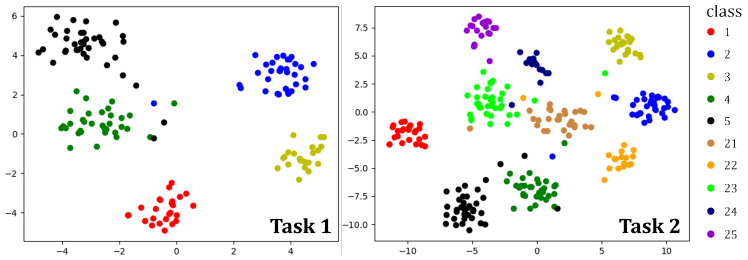

Fig. 4: The visualization of features in different tasks.

$\delta \in (0, 8)$. We report the results in Figure 2 and Figure 3. We can see that we selected the optimal values of the hyperparameter.

A.7 Visualization of Plasticity and Stability

To further demonstrate the effectiveness of C-ADA in acquiring new knowledge (plasticity) and in preventing the forgetting of old knowledge (stability) to solve the RFCL problem. we have visualized the t-SNE features of the output from C-ADA during the process of incremental learning, as depicted in Figure 4. It is evident from the visualization that after learning a new task, the features from the old and new classes are clustered together. This observation underscores the robustness of C-ADA maintaining a balance between plasticity and stability, thereby effectively solving the RFCL problem.

References

1. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022) 1
2. Ermis, B., Zappella, G., Wistuba, M., Rawal, A., Archambeau, C.: Memory efficient continual learning with transformers. Advances in Neural Information Processing Systems **35**, 10629–10642 (2022) 1

3. Gao, Q., Zhao, C., Ghanem, B., Zhang, J.: R-dfcil: Relation-guided representation learning for data-free class incremental learning. In: European Conference on Computer Vision. pp. 423–439. Springer (2022) 2, 3
4. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019) 1
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 1
6. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9374–9384 (2021) 2, 3
7. Smith, J.S., Karlinsky, L., et al.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: CVPR. pp. 11909–11919 (2023) 2
8. Tang, Y.M., et al.: When prompt-based incremental learning does not meet strong pretraining. In: CVPR (2023) 2
9. Yang, T., Huang, L., Luo, R.: Data-free class-incremental learning with implicit representation of prototypes. In: ECAI 2023, pp. 2866–2873. IOS Press (2023) 2, 3
10. Zhou, D.W., et al.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. arXiv (2023) 1, 2