Semantically-Shifted Incremental Adapter-Tuning is A Continual ViTransformer

Supplementary Material

1. More experiments on parameter sensitivity

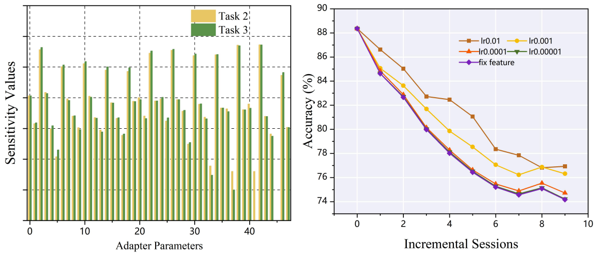

Figure 1. left: Parameter sensitivity of the different modules of task 2 and task 3. right: Incremental performance of varying degrees of update limits on important parameters.

We conduct parameter sensitivity experiments on ImageNetR, and the results are shown in Fig. 1. The left graph illustrates the parameter sensitivity of the second and third tasks. Similar to the trends observed in task one and two as shown in the main paper, the parameter sensitivity of different modules in task two and three also exhibits a high degree of similarity. Therefore, restricting parameter updates based on parameter sensitivity negatively impacts the learning of new categories. To investigate this phenomenon, we conduct experiments, as shown in the right graph. It can be seen that as the restriction on parameter updates in increases, the overall performance decreases.

2. More Ablation Experiments

Adapter dimension and layers insert: As shown in Tab. 1 and 2 left, we further conduct experiments on the specific position to insert the adapter module on the ImageNetR and ImageNetA datasets. We observe that the performance is progressively improved with the number of layers increased. Thus, we insert adapter modules in all 12 layers in our comparative experiments of various methods. We also conduct ablation experiments on the middle dimension in the adapter in Tab. 1 and 2 right. We observe that increasing the dimension has a positive effect on the performance of the model. Interestingly, setting the middle dimension to 32 did not result in a significant decrease in performance. On the other hand, setting it to 256 led to an improvement in performance but also quadrupled the number of parameters. To strike a balance between performance and the number of fine-tuning parameters, we set the middle dimension to 64.

Analysis of margin and scale: We conduct experiments on the hyper-parameter of the scale and margin in our cosine loss as shown in Tab. 5. We discover that appropriately increasing scale can enhance the performance of the model. For example, on the ImageNetR dataset, when the scale is set to 20, the average accuracy is 3.36% higher than when the value is 10. We also conduct experiments to analyze the influence of the margin. As shown in Tab. 6, we observe that different datasets have different appropriate margin values. For example, on the ImageNetR dataset, we set the margin to 0.0, and on the CUB200 dataset, we set it to 0.1.

Layers	Form	Acc	Num	#Param	Acc
1-3	parallel	80.53	32	0.60M	81.98
1-6	parallel	81.63	64	1.19M	82.09
1-12	parallel	**81.95**	256	4.87M	**82.38**

Table 1. Experimental results of the inserted layers of the adapter and the middle dimension.

Different PET methods in the CIL: The experiment results on CIFAR100 are in the main paper. In the supplementary material, we provide experiments with different PET methods on ImageNetR as shown in Tab. 7. We can observe that in the initial sessions, the performance of SSF surpasses that of the adapter. However, due to the tendency of SSF to overfit to the current session classes, there is a significant decline in subsequent incremental sessions and the adapter performs best in both the accuracy of the last session and average accuracy. We also provide experiments on ImageNetA as shown in Tab. 8. We can draw the same conclusion from ImageNetR.

Layers	Form	Acc	Num	#Param	Acc
1-3	parallel	65.25	32	0.60M	66.49
1-6	parallel	65.81	64	1.19M	66.85
1-12	parallel	**66.67**	256	4.87M	**67.54**

Table 2. Experimental results of the inserted layers of the adapter and the middle dimension.

Unified classifier retraining vs. Separate local classifier: The experiment results on ImageNetA are in the main paper. Here we show the accuracy of each session of three different seeds on ImageNetA as shown in Tab. 9. It can be seen that retraining the classifier can improve the performance by 2% to 3% on three seeds, effectively improving the performance of the classifier. Furthermore, classifier retraining with semantic shift estimation can further improve performance by 2% to 3%. We also show the results on CUB200 as shown in Tab. 10. The same trend is shown in this dataset. CA can significantly improve the performance, and SSCA can align the prototype and further improve performance. The results on ImageNetR are shown in Tab. 11.

Different pre-trained models. We experiment with pre-trained models (PTMs) with different generalization abilities on ImageNetR and ImageNetA datasets shown in Tab. 3. It can be observed that our method generalizes well to various PTMs. The large-based ViT model can get better performance.

Different tuning methods with SSCA. We incorporate classifier alignment with semantic shift estimation into SSF and prompt tuning shown in Tab. 4. It can be seen that both the performance of prompt-based and SSF tuning approaches show significant improvement. However, our proposed method still outperforms them by a large margin. The results further verify the effectiveness of our proposed method.

Pre-trained Model	ImageNetR		ImageNetA	
	$\mathcal{A}_{Last}$ ↑	$\mathcal{A}_{Avg}$ ↑	$\mathcal{A}_{Last}$ ↑	$\mathcal{A}_{Avg}$ ↑
ViT-base 1K	$80.15_{\pm 0.41}$	$83.87_{\pm 0.26}$	$64.88_{\pm 1.11}$	$72.68_{\pm 1.72}$
ViT-base 21k	$79.38_{\pm 0.59}$	$83.63_{\pm 0.43}$	$62.43_{\pm 1.63}$	$70.83_{\pm 1.63}$
ViT-large 21k	$83.62_{\pm 0.41}$	$86.70_{\pm 0.69}$	$68.38_{\pm 2.25}$	$74.85_{\pm 1.93}$

Table 3. Results on different pre-trained models on ImageNetR/A.

Method	ImageNetR		ImageNetA	
	$\mathcal{A}_{Last}$ ↑	$\mathcal{A}_{Avg}$ ↑	$\mathcal{A}_{Last}$ ↑	$\mathcal{A}_{Avg}$ ↑
SSF	$71.84_{\pm 0.33}$	$79.98_{\pm 0.79}$	$52.11_{\pm 0.64}$	$62.34_{\pm 1.33}$
+SSCA	$75.01_{\pm 0.31}$	$82.09_{\pm 0.41}$	$58.94_{\pm 1.09}$	$67.94_{\pm 1.06}$
VPT-deep	$38.49_{\pm 0.13}$	$50.34_{\pm 1.93}$	$37.39_{\pm 22.03}$	$46.55_{\pm 16.69}$
+SSCA	$56.11_{\pm 3.25}$	$61.11_{\pm 1.71}$	$47.83_{\pm 18.75}$	$55.67_{\pm 14.92}$
VPT-shallow	$58.79_{\pm 1.07}$	$69.23_{\pm 4.06}$	$48.34_{\pm 0.99}$	$56.96_{\pm 3.45}$
+SSCA	$68.25_{\pm 2.50}$	$72.40_{\pm 2.23}$	$54.49_{\pm 0.76}$	$62.26_{\pm 2.54}$

Table 4. Results for different PET methods on ImageNetR/A.

3. More Implementation Details

To mitigate the impact of randomness in the experiments, we selected three different seeds (1993,1996, and 1997) to conduct experiments separately and calculate the average and variance. In experiments involving different PET methods, we fine-tuned the parameters inserted into the network without unified classifier retraining. For experiments of adapter dimension and layers insert, we conduct experiments on the ImageNetR dataset and we set the loss margin to 0.0 and the scale to 20. In the analysis of margin experiments, we set the scale to 20 for all datasets. In the analysis of scale experiments, we set the scale to 0.0 for all datasets.

References

[1] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3

[2] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[3] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 3

Scale	ImageNetR		ImageNetA		CIFAR100		CUB200	
	Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑
s=10	73.80	80.84	56.95	66.88	89.49	93.61	86.39	91.61
s=15	77.83	83.36	59.84	69.25	91.02	94.53	88.46	92.57
s=20	**79.55**	**84.20**	**60.76**	**69.62**	**91.62**	**94.75**	88.51	**92.83**
s=30	78.90	83.30	60.50	68.73	91.21	94.51	**88.60**	92.64

Table 5. Experimental results of the influence of scale in cosine loss on different datasets.

Margin	ImageNetR		ImageNetA		CIFAR100		CUB200	
	Last↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑	Last ↑	Avg ↑
m = 0	**79.55**	84.20	60.76	69.62	**91.62**	**94.75**	88.13	92.33
m = 0.1	78.38	**84.25**	**63.00**	72.13	91.60	94.71	**88.46**	**92.57**
m = 0.2	76.48	82.94	62.74	**73.18**	89.71	93.28	88.38	92.56
m = 0.3	73.90	80.76	62.61	72.46	86.83	91.45	87.57	91.89

Table 6. Experimental results of the influence of margin in cosine loss on different datasets.

PET Method	Params	Ses.1	Ses.2	Ses.3	Ses.4	Ses.5	Ses.6	Ses.7	Ses.8	Ses.9	Ses.10	Avg↑
SSF [3]	0.2M	94.75	89.28	86.11	82.61	80.19	78.67	77.27	75.4	75.04	72.78	81.21
VPT-deep [2]	0.046M	87.23	64.19	59.16	40.27	39.4	36.6	31.51	33.32	33.79	31.62	45.71
VPT-shallow [2]	0.004M	81.86	73.20	68.81	66.85	64.89	63.81	62.84	62.21	61.35	58.97	66.48
Adapter [1]	1.19M	91.87	88.42	86.51	84.43	82.75	81.51	80.99	80.62	79.75	78.28	83.51

Table 7. Experimental results for baselines with different parameter efficient tuning methods on ImageNetR. We report the overall performance of each session and the average performance.

PET Method	Params	Ses.1	Ses.2	Ses.3	Ses.4	Ses.5	Ses.6	Ses.7	Ses.8	Ses.9	Ses.10	Avg↑
SSF [3]	0.2M	82.86	76.11	67.44	65.62	61.47	58.97	54.33	52.32	51.04	51.28	62.14
VPT-deep [2]	0.046M	61.14	31.67	15.55	15.62	10.11	6.05	4.82	4.17	3.37	3.42	15.92
VPT-shallow [2]	0.004M	80.00	70.00	64.92	60.32	56.84	54.26	52.63	51.52	49.68	48.26	58.84
Adapter	1.19M	82.86	74.72	71.43	67.62	64.68	62.15	59.14	56.89	55.20	55.50	65.02

Table 8. Experimental results for baselines with different parameter efficient tuning methods on ImageNetA. We report the overall performance of each session and the average performance.

Seed	Method	Ses.1	Ses.2	Ses.3	Ses.4	Ses.5	Ses.6	Ses.7	Ses.8	Ses.9	Ses.10	Avg ↑
1993	w/o CA	85.71	81.11	75.84	72.92	68.85	64.21	60.48	60.18	59.07	58.66	68.70
	w/ CA	85.71	81.94	77.10	73.07	71.22	68.00	64.14	63.22	60.14	59.91	70.45
	w/ SSCA	85.71	83.61	78.57	76.36	73.84	72.00	66.73	65.06	62.37	62.15	72.64
1996	w/o CA	81.99	77.34	71.36	64.74	66.14	63.83	62.54	60.25	60.09	58.39	66.67
	w/ CA	81.99	78.06	74.57	67.91	67.15	64.57	64.59	62.62	61.92	62.08	68.55
	w/ SSCA	81.99	78.06	74.81	69.78	71.18	67.9	66.43	65	64.2	64.19	70.35
1997	w/o CA	80.29	74.91	69.41	65.23	63.03	61.49	56.89	57.63	57.62	56.68	64.32
	w/ CA	80.29	79.09	72	70.22	66.21	64.41	61.46	59.46	60.03	57.34	67.05
	w/ SSCA	80.29	80.14	74.35	71.08	69.66	67.12	65.15	63.73	62.44	60.96	69.49

Table 9. Ablation results for unified classifier training and semantic shift estimation on ImageNetA. We report the overall performance of each session and the average performance.

Seed	Method	Ses.1	Ses.2	Ses.3	Ses.4	Ses.5	Ses.6	Ses.7	Ses.8	Ses.9	Ses.10	Avg ↑
1993	w/o CA	99.19	95.37	91.37	87.81	85.86	82.49	82.15	80.96	79.09	78.63	86.29
	w/ CA	99.19	98.24	93.45	91.62	91.33	88.42	87.69	86.78	86.37	85.50	90.86
	w/ SSCA	99.19	98.24	94.64	93.14	93.10	91.45	90.68	90.59	89.71	88.80	92.95
1996	w/o CA	100.00	94.79	91.09	90.02	88.34	85.92	85.59	82.74	80.87	79.05	87.84
	w/ CA	100.00	95.83	94.60	92.60	92.00	90.83	90.56	86.84	85.40	85.58	91.42
	w/ SSCA	100.00	96.25	96.06	94.96	94.30	93.81	93.32	91.26	90.22	89.10	93.93
1997	w/o CA	96.55	97.20	91.9	87.87	84.92	84.28	82.15	82.1	82.69	78.92	86.86
	w/ CA	96.55	97.90	95.8	90.67	90.12	90.29	87.96	87.43	86.98	85.88	90.96
	w/ SSCA	96.55	97.67	95.8	92.02	91.25	91.30	89.56	89.5	89.39	88.34	92.14

Table 10. Ablation results for unified classifier training and semantic shift estimation on CUB200. We report the overall performance of each session and the average performance.

Seed	Method	Ses.1	Ses.2	Ses.3	Ses.4	Ses.5	Ses.6	Ses.7	Ses.8	Ses.9	Ses.10	Avg ↑
1993	w/o CA	91.73	88.64	86.14	84.59	82.08	81.32	80.68	80.52	79.67	78.47	83.38
	w/ CA	91.73	88.12	86.20	84.51	82.08	81.26	81.32	81.00	79.38	78.02	83.36
	w/ SSCA	91.73	88.57	86.72	84.95	82.91	82.19	82.32	81.63	80.43	79.58	84.10
1996	w/o CA	89.55	86.63	85.27	82.45	81.64	80.36	79.56	78.33	78.34	77.40	81.95
	w/ CA	89.55	87.59	86.37	83.58	82.15	81.12	79.85	78.81	78.51	77.67	82.52
	w/ SSCA	89.55	87.11	86.65	84.04	83.34	82.00	81.51	79.87	79.73	78.72	83.25
1997	w/o CA	91.20	87.90	85.17	82.96	80.57	80.59	79.35	78.14	77.85	77.68	82.14
	w/ CA	91.20	88.73	85.52	83.47	81.88	81.02	79.70	79.06	78.42	78.40	82.74
	w/ SSCA	91.20	89.00	86.16	83.68	82.40	81.75	80.97	80.63	79.73	79.85	83.54

Table 11. Ablation results for unified classifier training and semantic shift estimation on ImageNetR. We report the overall performance of each session and the average performance.