Consistent Prompting for Rehearsal-Free Continual Learning

Supplementary Material

1. Smooth Regularization analysis

The impact of τ on the adaptative entropy loss (Eq. (3)) and smooth regularization process are discussed in detail. Suppose that there are n_i classes seen within task i. $g_j = \sigma_j(\ell_i/\tau)$ and $\rho_j = \sigma_j(\ell_i)$ refer to the probabilities of the jth class based on different logits. Therefore, Eq. (3) can be rewrite as,

$$\begin{aligned}\mathcal{L}_e(i) &= - < \sigma(\ell_i/\tau),\ \log(\sigma(\ell_i)) > \\ &= -\sum_{j=1}^{n_i} g_j \log \rho_j.\end{aligned} \tag{a}$$

The gradient of $\mathcal{L}_e(i)$ with respect to ρ_j can be calculated,

$$\begin{aligned}\frac{\partial \mathcal{L}_e(i)}{\partial \rho_j} &= \frac{\partial}{\partial \rho_j}(-\sum_{j=1}^{n_i} g_j \log \rho_j) \\ &= \frac{\partial}{\partial \rho_j}(-\sum_{j=1}^{n_i-1} g_j \log \rho_j - g_{n_i} \log \rho_{n_i}) \\ &= \frac{\partial}{\partial \rho_j}(-\sum_{j=1}^{n_i-1} g_j \log \rho_j - g_{n_i} \log (1 - \sum_{j=1}^{n_i-1} \rho_j)) \\ &= -\frac{g_j}{\rho_j} + \frac{g_{n_i}}{1 - \sum_{j=1}^{n_i-1} \rho_j} \\ &= -\frac{g_j}{\rho_j} + \frac{g_{n_i}}{\rho_{n_i}},\end{aligned} \tag{b}$$

where $\rho_n = 1 - \sum_{j=1}^{n_i-1} \rho_j$. The impact of this gradient on ρ_j is,

$$\rho_j := \rho_j - \frac{\partial \mathcal{L}_e(i)}{\partial \rho_j}. \tag{c}$$

Given that the gradient from g_j, i.e., $\sigma(\ell_j/\tau)$, are blocked and does not flow back, $g_j = \rho_j$ and $g_{n_i} = \rho_{n_i}$ when $\tau = 1$ and the corresponding gradient $\frac{\partial \mathcal{L}_e(i)}{\partial \rho_j}$ are all zeros. Therefore, the smooth regularization will be turned off. More importantly, the proposed loss of Eq. (3) has the desired smoothing effect by setting $\tau > 1$. Specifically, when ρ_j reaches a large value, e.g. close to 1, $\frac{g_j}{\rho_j}$ will be smaller than $\frac{g_{n_i}}{\rho_{n_i}}$, resulting in positive gradients to lower ρ_j down. In contrast, if ρ_j reaches a small value, for example close to 0, $\frac{g_j}{\rho_j}$ will be greater than $\frac{g_{n_i}}{\rho_{n_i}}$, resulting in negative gradients to lift ρ_j.

Intuitive alternatives to Smooth Regularizations Two smoothing techniques are proposed and compared with ours. The first technique provides an evenly distributed label (EDL), $\frac{1}{n_i}$ on each entry, as the supervision. Its learning objective becomes,

$$\mathcal{L}_e(i) = - < \frac{1}{n_i},\ \log(\sigma(\ell_i)) > . \tag{d}$$

Another smoothing baseline involves the one-hot label δ by randomly activating a category entry. This smoothing technique is denoted random one-hot, ROH, with cross-entropy objective,

$$\mathcal{L}_e(i) = - < \delta,\ \log(\sigma(\ell_i)) > . \tag{e}$$

Comparisons between the proposed smooth regularization and the alternative ones are shown in Table A. Our approach outperforms the two baseline methods by a significant margin.

Table A. Results of different regularization methods on 10-task continual learning of Split StanfordCars.

	Last-acc ↑	Avg-acc ↑
EDL	39.73	50.93
ROH	40.99	51.59
Ours	66.77	76.81

2. Recent relevant methods

SLCA [2] and HiDe-Prompt [1] handle the classifier inconsistency with a post-hoc alignment technique to boost performance. Both methods involve storing the mean and covariance of each class feature and constructing Gaussian distributions for each class. Subsequently, unified classifiers are retrained by sampling features generated from these Gaussian distributions. However, saving the covariance of each class feature incurs significant memory overhead, particularly considering the sensitivity of continual learning to memory usage. Notably, our proposed classifier consistency learning approach only introduces regularization without requiring additional memory overhead.

3. Analysis of the FF score

According to the FF score computation

$$FF_t = \frac{1}{t-1}\sum_{j=1}^{t-1}\{\max_{i\in\{1,\ldots,t-1\}}(a_{i,j}) - a_{t,j}\}, \forall j < t, \tag{f}$$

all $a_{t,j}$ over the 20 tasks in Table 1 can be depicted as in Figure A , where CPrompt and L2P (with the best FF score) are compared. The $\max_{i\in\{1,\ldots,t-1\}}(a_{i,j})$ of

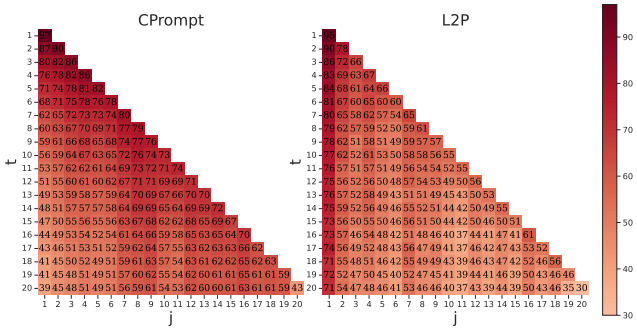

Figure A. Detailed illustration of FF score. All $a_{t,j}$ values are recorded and presented with a heat map. The deeper the red, the higher the value. Best viewed in colors and zoom-in for the values.

CPrompt are larger than those of L2P, as the diagonal area of CPrompt shows deeper red than that of L2P, resulting in larger differences and thus higher FF of CPrompt. Moreover, CPrompt achieves superior overall performance and stability-plasticity trade-off, as Figure A suggests.

4. More results with our approach

As shown in Table B, PCL and CCL consistently improve various backbones with clear margins.

Table B. Continual learning results of the Split StanfordCars 10-task. The improvements over the backbone are shown in red. L: Last-acc, A: Avg-acc.

Method	CODAPrompt		L2P	
	L	A	L	A
Backbone	62.24	73.28	60.39	71.92
+CCL	65.12+2.88	75.37+2.09	63.69+3.30	72.53+0.61
+PCL	63.49+1.25	74.21+0.93	61.82+1.43	72.06+0.14
+BOTH	65.31+3.07	75.81+2.53	63.71+3.32	72.71+0.79

5. Resource consumption

Memory usage of a method is evaluated by the number of trainable parameters. Its computational requirement is evaluated by FLOPs. As shown in Table C, the proposed method consumes both resources at relatively low levels.

Table C. Different methods' computation and memory consumption on Split StanfordCars 10-task.

Method	Ours	ESN	L2P	Dual	CODA
Params ($\times 10^6$)	0.92	30.49	0.78	0.44	89.79
FLOPs ($\times 10^9$)	23.62	23.97	23.62	23.62	33.77

References

[1] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[2] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023. 1