CAPrompt: Cyclic Prompt Aggregation for Pre-Trained Model Based Class Incremental Learning

Qiwei Li, Jiahuan Zhou*

Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China
{lqw, jiahuanzhou}@pku.edu.cn

Abstract

Recently, prompt tuning methods for pre-trained models have demonstrated promising performance in Class Incremental Learning (CIL). These methods typically involve learning task-specific prompts and predicting the task ID to select the appropriate prompts for inference. However, inaccurate task ID predictions can cause severe inconsistencies between the prompts used during training and inference, leading to knowledge forgetting and performance degradation. Additionally, existing prompt tuning methods rely solely on the pre-trained model to predict task IDs, without fully leveraging the knowledge embedded in the learned prompt parameters, resulting in inferior prediction performance. To address these issues, we propose a novel Cyclic Prompt Aggregation (CAPrompt) method that eliminates the dependency on task ID prediction by cyclically aggregating the knowledge from different prompts. Specifically, rather than predicting task IDs, we introduce an innovative prompt aggregation strategy during both training and inference to overcome prompt inconsistency by utilizing a weighted sum of different prompts. Thorough theoretical analysis demonstrates that under concave conditions, the aggregated prompt achieves lower error compared to selecting a single task-specific prompt. Consequently, we incorporate a concave constraint and a linear constraint to guide prompt learning, ensuring compliance with the concave condition requirement. Furthermore, to fully exploit the prompts and achieve more accurate prompt weights, we develop a cyclic weight prediction strategy. This strategy begins with equal weights for each task and automatically adjusts them to more appropriate values in a cyclical manner. Experiments on various datasets demonstrate that our proposed CAPrompt outperforms state-of-the-art methods by 2%-3%.

Code — https://github.com/zhoujiahuan1991/AAAI2025-CAPrompt

Introduction

In recent years, the field of computer vision has seen remarkable advancements, particularly in the domain of image classification (LeCun et al. 1989; Deng et al. 2009; Dosovitskiy et al. 2020). However, traditional learning models typically assume that the entire training dataset is available from the onset (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016). This assumption leads to the challenge when models are required to adapt to new data incrementally without forgetting previously learned information, a phenomenon commonly referred to as *catastrophic forgetting* (French 1999). *Class Incremental Learning* (CIL) (Zhou et al. 2024b; Wang et al. 2024b) aims to address this issue by enabling models to learn continuously from a stream of data, adapting to new tasks while retaining past knowledge.

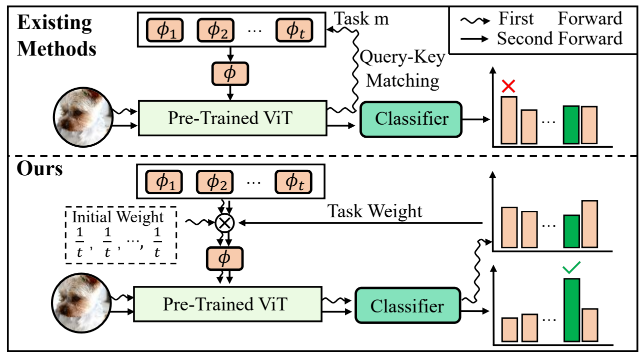

Figure 1: Most existing prompt-based methods (Dualprompt) predict task ID during inference which may cause inconsistency between the prompts during training and inference. In contrast, we propose a prompt aggregation strategy to eliminate the requirement to predict task ID. Moreover, a cyclic prompt weight strategy is proposed to adjust the weights of different prompts.

The primary challenge of CIL is balancing the acquisition of new knowledge and the retention of previous knowledge (Zhou et al. 2024b). A promising solution lies in leveraging pre-trained models, which are trained on large, diverse datasets to capture rich and generalized feature representations. Recent methods aim to utilize Parameter-Efficient Fine-Tuning (PEFT) techniques, such as prompt learning (Wang et al. 2022d,c), to acquire new knowledge while keeping the pre-trained model frozen, thereby mitigating forgetting. During training, these methods not only learn a set of task-specific prompt parameters but also feed the input sample into the pre-trained model without prompts to learn a set of keys for each task, which are used for task se-

*Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

lection by query-key matching during inference. As demonstrated in Fig.1, in the inference stage, due to the absence of task ID information, these methods need to infer the task ID based on the learned keys at first. In the subsequent inference stage, the predicted task ID is used to select the relevant task-specific prompt for the pre-trained model. However, the predicted task ID for certain samples may not match their ground truth ID, resulting in the adoption of prompts associated with incorrect tasks during inference. This inconsistency between the prompts used during training and inference inevitably leads to severe knowledge forgetting, thereby decreasing final performance (Gao, Cen, and Chang 2024). Furthermore, existing methods rely solely on the pre-trained model to predict task ID without fully leveraging the knowledge embedded in the learned prompt parameters, exacerbating the degradation in performance.

To address these challenges, we propose a novel **C**yclic **Prompt A**ggregation (**CAPrompt**) method for CIL with pre-trained models. As illustrated in Fig.1, instead of predicting task IDs to select one task-specific prompt, we estimate the probability of each class and aggregate these probabilities into the probability of each task. These task probabilities are then used as weights for prompts at different tasks to generate the aggregated prompt. Unlike a few existing methods that employ similar weighted sums of prompts (Smith et al. 2023; Roy et al. 2024), we provide a theoretical justification that the prediction error of the aggregated prompt is lower than using one task-specific prompt parameter in a single stage, under the condition that the network prediction is concave with respect to the prompt parameters for a given image. To ensure the network satisfies this concave condition, we introduce a concave constraint and a linear constraint. Furthermore, to fully utilize the knowledge embedded in the prompts to guide the pre-trained models, we develop a cyclic prompt weight prediction strategy. Instead of merely predicting the prompt weight without the guidance of prompts, we initialize with equal weights for each task and cyclically adjust the prompt weights to more appropriate values. Additionally, beyond the conventional two-stage paradigm, our method can be performed cyclically multiple times, further boosting performance.

In summary, the contributions of this paper are four-fold: (1) A new Prompt Aggregation strategy is proposed for both training and inference, eliminating the need for task ID prediction thus mitigating the inconsistencies of prompts caused by task prediction errors. (2) Comprehensive theoretical analysis is provided to demonstrate that our aggregated prompt achieves lower prediction error compared to using a single task-specific prompt under the concave condition. The concave and linear constraints are proposed to facilitate the model to satisfy this condition. (3) A cyclic prompt weight prediction strategy is proposed to cyclically adjust the weights of prompts to more accurate ones, further improving performance. (4) Extensive experiments on various benchmarks demonstrate that our CAPrompt outperforms state-of-the-art approaches by 2%-3%.

Related Work

Class Incremental Learning

Current Class Incremental Learning methods can be broadly categorized into *rehearsal-based*, *regularization-based*, and *architecture-based* methods. Rehearsal-based approaches (Prabhu, Torr, and Dokania 2020; Liu, Schiele, and Sun 2021; Luo et al. 2023) selected and stored representative samples from earlier classes to explicitly preserve knowledge. During the training of later classes, they replayed these stored samples to mitigate forgetting. Moreover, regularization-based methods (Kirkpatrick et al. 2017; Li and Hoiem 2017; Smith et al. 2021; Li et al. 2024; Li, Peng, and Zhou 2024b,a; Xu, Zou, and Zhou 2024) stabilized model parameters and feature adjustments to address forgetting. Some focus on preserving knowledge by maintaining consistency of certain metrics (e.g. logits or feature similarity), while others directly restricted changes to important model parameters. The architecture-based models (Wang et al. 2022a,b; Zhou et al. 2022; Hu et al. 2023; Xu et al. 2024) dynamically expanded network structures to adapt to the evolving data stream. These methods typically froze previously learned parameters and initialize new parameters to learn the knowledge of new classes.

Pre-Trained Model Based CIL

Recently pre-trained model based CIL has attracted rising attention due to the promising results (Yu et al. 2024). Most Pre-Trained Model-based CIL methods utilize the Parameter-Efficient Fine-Tuning (PEFT) mechanism to adapt the model efficiently while keeping the Pre-Trained model frozen. For example, SSIAT (Tan et al. 2024) and EASE (Zhou et al. 2024a) adapted the model using adapter (Chen et al. 2022) and approximate the feature of previous classes to mitigate forgetting. Additionally, a substantial number of methods leverage prompts. L2P (Wang et al. 2022d) designed a prompt pool for incremental learning, selecting instance-specific prompts during training and inference. Most methods (Wang et al. 2022c, 2023; Qiao et al. 2023; Wang et al. 2024a; Liu, Peng, and Zhou 2024) developed task-specific prompts, keeping prompts for other tasks frozen during training. Instead of directly optimizing prompt parameters, DAP (Jung et al. 2023) designed prompt generators to generate more instance-specific information in prompts. However, these methods require knowledge of the task information for inference samples, necessitating task ID inference first to select the relevant prompt. And errors in task prediction can lead to inconsistencies between training and inference, reducing their performance. A few methods (Smith et al. 2023; Roy et al. 2024) proposed a weighted sum of prompts that can mitigate this problem. However, they did not thoroughly explore the mechanism of weighted prompt, which limits the effectiveness of their weighted prompt strategies. In this paper, we propose a prompt aggregation strategy with concave and linear constraints which guarantee the prediction error of aggregated prompt is lower than task-specific prompt. A cyclic prompt weight strategy is also proposed to fully utilize the learned prompts to adjust the prompt weight, which further improves performance.

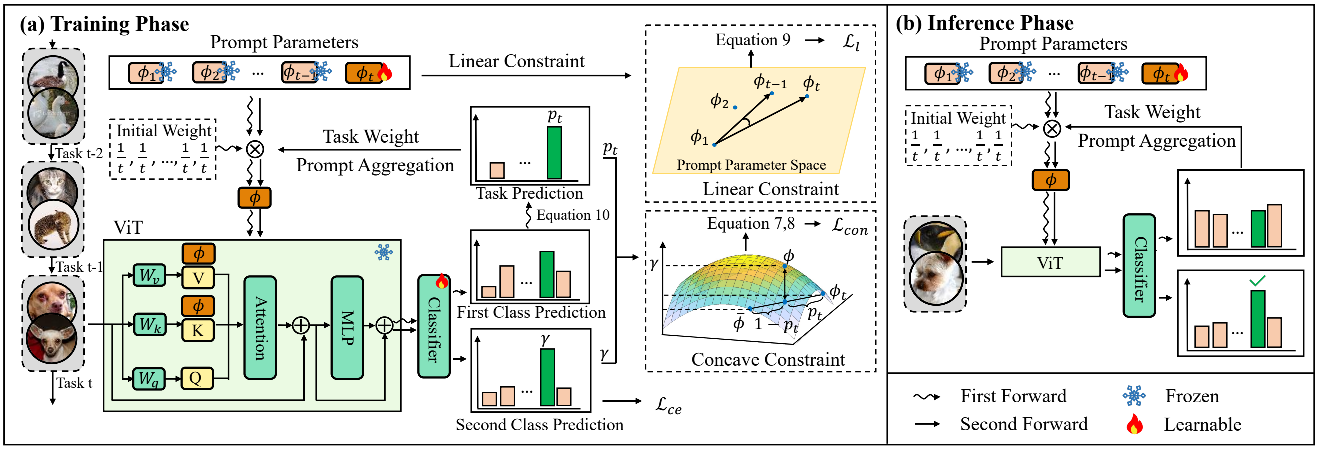

Figure 2: The overall pipeline of our proposed CAPrompt. To overcome the inconsistency of prompts between the training phase (a) and inference phase (b), a prompt aggregation strategy is proposed. The concave and linear constraints are proposed to guarantee the prediction error of the aggregated prompt is the lower bound of using one task-specific prompt. Then, to make full use of prompts in predicting weight for prompt aggregation, we propose a cyclic prompt weight that we initiate the prompt with equal weight and cyclically predict the prompt weight. This strategy can be conducted cyclically many times to further improve performance.

Cyclic Prompt Aggregation

Problem Formulation

Class Incremental Learning involves learning a model from a data stream. For a data stream with T tasks, $\mathcal{D} = \{D_t\}_{t=1}^T$, each dataset $D_t = \{X_t, Y_t\}$ consists of input data set $X_t = \{x_{t,j}\}_{j=1}^{n_t}$ and a label set $Y_t = \{y_{t,j} \in \mathcal{C}_t\}_{j=1}^{n_t}$, where n_t is the number of data in task t, $x_{t,j}$ represents the j-th image and $y_{t,j}$ represents the label. $\mathcal{C}_t$ is the label set and labels of different tasks are disjoint, that is $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset (i \neq j)$. In task t, the model learns a mapping function $f_t : \mathbb{R}^{h\times w\times 3} \to \mathbb{R}^{l_t}$ for all seen classes, where $l_t = \sum_{j=1}^{t} |\mathcal{C}_j|$ is the number of classes that have been learned up to task t. In the pre-trained model based CIL, f_t typically consists of a frozen Pre-Trained model θ (e.g. Vision Transformer), the classification head W, and learnable parameters ϕ_t (e.g. prompts). Thus, for an input image x, the prediction with prompt ϕ_t is $f_t(x, W)$, which can also be formed as $W^\top f(x, \phi_t)$. The prediction of solely using the pre-trained model is $W^\top f(x)$.

Prompt Aggregation

As mentioned above, existing methods need to decide which task-specific prompt to use during inference, and errors in task prediction can lead to inconsistencies between training and inference, resulting in knowledge forgetting. To mitigate this problem, we propose the Prompt Aggregation strategy which aggregates the knowledge of prompts for different tasks instead of selecting a single task-specific prompt.

Firstly, during the training of task t, given an image x with label y, instead of only utilizing the task-specific prompt ϕ_t, we calculate the task similarity for task i using query $f(x)$,

$$p_i = \sum_{m\in\mathcal{C}_i} \mathrm{Softmax}(W^\top f(x))[m], \tag{1}$$

where W consists of the maintained key feature for each seen class, $\mathrm{Softmax}$ represents softmax operation. Then the aggregated prompt can be calculated by:

$$\phi = \sum_{i=1}^{t-1} p_i \cdot \mathrm{stop}(\phi_i) + p_t \cdot \phi_t. \tag{2}$$

Here, we use stop gradient operation, stop, for the prompt of previous tasks to prevent the interference of previous prompt parameters. Then the optimization function with cross-entropy loss is:

$$\mathcal{L}_{ce} = CE(W^\top f(x, \phi), y). \tag{3}$$

During inference, in the first stage, we obtain the aggregated prompt according to Eq.1 and Eq.2. Then in the second stage, the prediction of x is given by $\mathrm{argmax}_m(W^\top f(x, \phi)[m])$. This prompt aggregation strategy not only ensures the consistency of prompts between training and inference but also satisfies the following theorem.

To simplify the notation, we denote the prediction of label y using prompt ϕ, $\mathrm{Softmax}(W^\top f(x, \phi))[y]$ as $g(x, y, \phi)$, where W represents the classification head for all classes.

Theorem 1: *In pre-trained based CIL, for dataset $D = \{X, Y\}$, the expected error (E_1) of using aggregated prompt, with the weight of each prompt being the sum of probability of classes of each task, is lower than the expected error (E_2) of using task-specific prompts, if the prediction of each class is concave to the combination of prompts for different tasks.*

Proof: The expected error of the aggregated prompt is:

$$E_1 = \mathbb{E}_x[-\log g(x, y, \sum_{i=1}^{t} p_i \cdot \phi_i)]. \tag{4}$$

The expected error using task-specific prompts is:

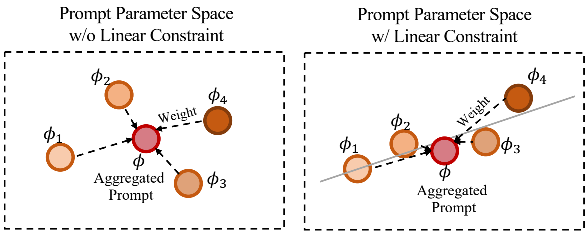

Figure 3: Motivation of Linear Constraint.

$$E_2 = \mathbb{E}_x[-\log \sum_{i=1}^{t} p_i \cdot g(x, y, \phi_i)], \tag{5}$$

where p_i is the task similarity in Eq.1, represents the probability of task i being the most similar task. When the prediction of the label y is concave in the prompt parameter space, according to Jensen Inequality:

$$g(x, y, \sum_{i=1}^{t} p_i \cdot \phi_i) \geq \sum_{i=1}^{t} p_i \cdot g(x, y, \phi_i), \tag{6}$$

then $E_1 \leq E_2$. Finish the proof (detailed in Supplementary).

This theorem demonstrates that the aggregated prompt yields a lower prediction error than using one task-specific prompt under the concave condition. Thus, to satisfy this condition, we propose the following **Concave Constraint** and **Linear Constraint**.

Concave Constraint. An intuitive way to meet with the concave condition is to directly constrain the prediction with different prompts during training. Thus for an image x with label h, the Concave Constraint for task t is defined as:

$$\begin{aligned} \delta = p_t \cdot g(x, y, \phi_t) + (1 - p_t) \cdot g(x, y, \sum_{i=1}^{t-1} \frac{p_i}{1 - p_t} \phi_i) \\ - g(x, y, \sum_{i=1}^{t} p_i \cdot \phi_i), \end{aligned} \tag{7}$$

$$\mathcal{L}_{con} = \max(\delta, 0). \tag{8}$$

By employing the concave loss in each task, we explicitly constrain the prompt to satisfy the concave condition.

Linear Constraint. Though we propose a concave constraint to ensure the model satisfies the concave condition in Theorem 1, this condition cannot be exactly met due to the large number of parameters. Moreover, the concave condition in Theorem.1 focuses on the combination of learned prompts for different tasks rather than the whole prompt parameters space. Thus we tend to restrict the learning of prompts to make the concave condition easier to meet. As illustrated in Fig.3, due to the aggregated prompt being the linear combination of different prompts, we propose a linear constraint to ensure that the linear combination of prompts is also close to the learned prompts. Specifically, the linear constraint forces the prompts of different tasks to lie in a linear direction by maximizing the cosine similarity of the prompt direction relative to the prompt of the first task. The linear constraint is formulated as:

$$\mathcal{L}_l = 1 - \mathrm{sim}(\phi_t - \phi_1, \phi_{t-1} - \phi_1), \tag{9}$$

where sim represents the cosine similarity.

In conclusion, by employing prompt aggregation in both training and inference, we eliminate the task prediction and ensure the consistency of prompts, thus mitigating knowledge forgetting. Additionally, the concave and linear constraints are designed to ensure that the aggregated prompt yields a lower prediction error than using a single task-specific prompt.

Algorithm 1: Cyclic Prompt Aggregation algorithm

Training:
Input: Data stream $\mathcal{D} = \{D_t\}_{t=1}^{T}$, a pre-trained ViT $f(\cdot)$.
Output: Prompts $\phi_1, ..., \phi_T$ and classification head W.

for t in $1 : T$ **do**
  Set initial weight $p_1^1 = ... = p_t^1 = \frac{1}{t}$.
  Get aggregated prompt ϕ^1 by Eq.2 with $p_1^1, ..., p_t^1$.
  Get task probability $p_1^2, .., p_t^2$ by Eq.10 with ϕ^1.
  Get aggregated prompt ϕ^2 by Eq.2 with $p_1^2, ..., p_t^2$.
  Compute the loss $\mathcal{L}_{ce} + \mathcal{L}_{con} + \mathcal{L}_l$ with ϕ^2 and update parameters.
end for

Inference after training on task t:
Input: Input image x, a pre-trained ViT $f(\cdot)$, prompts $\phi_1, ..., \phi_t$, classification head W, number of cycles num.
Output: Class number m.

Let $n = 1$.
Set initial weight $p_1^n = ... = p_t^n = \frac{1}{t}$.
for n in $1 : num$ **do**
  Get aggregated prompt ϕ^n by Eq.2 with $p_1^n, .., p_t^n$.
  Get task probability $p_1^{n+1}, .., p_t^{n+1}$ by Eq.10 with ϕ^n.
end for
$m = \mathrm{argmax}_m(W^\top f(x, \phi^{num})[m])$.

Cyclic Prompt Weight

Though the proposed Prompt Aggregation enhances the consistency of prompts, the aggregation weight of prompts for each task is obtained by the pre-trained model without prompts which neglects the knowledge of prompts. To fully utilize the learned prompts, we propose a cyclic prompt weight prediction strategy. In the first stage during training and inference, we get the aggregated prompts ϕ^1 by Eq.2 with equal weight for each prompt, $p_1^1 = p_2^1 = ... = p_t^1 = \frac{1}{t}$, and get the class probability prediction. Then the Eq.1 can be formed as follows to calculate the new task probability for cycle $n + 1$ with aggregated prompt n:

$$p_i^{(n+1)} = \sum_{m \in \mathcal{C}_i} \mathrm{Softmax}(W^\top f(x, \phi^n))[m]. \tag{10}$$

This task probability serves as the weight to calculate the new aggregated prompts ϕ^2 by Eq.2. This updated prompt ϕ^2 is then used to prompt the model in the second stage.

Methods	Publication	CIFAR-100		ImageNet-R		CUB200	
		ACC	AF	ACC	AF	ACC	AF
L2P	CVPR'22	83.06 ± 0.40	5.95 ± 0.47	63.22 ± 0.34	7.05 ± 0.37	70.88 ± 0.03	6.04 ± 0.39
DualPrompt	ECCV'22	85.11 ± 0.29	5.74 ± 0.46	69.14 ± 0.16	4.55 ± 0.01	76.54 ± 0.15	5.67 ± 0.20
CODA-Prompt	CVPR'23	86.76 ± 0.22	6.36 ± 0.28	73.71 ± 0.25	5.11 ± 0.65	73.71 ± 0.92	7.49 ± 0.20
HiDe-Prompt	NeurIPS'23	92.64 ± 0.22	1.92 ± 0.23	75.66 ± 0.21	**2.88 ± 0.19**	86.84 ± 0.14	2.07 ± 0.05
EvoPrompt	AAAI'24	87.57 ± 0.40	5.50 ± 0.44	76.49 ± 0.27	3.57 ± 0.13	79.88 ± 0.31	9.40 ± 0.81
ConvPrompt	CVPR'24	88.86 ± 0.21	3.37 ± 0.24	77.94 ± 0.06	3.43 ± 0.16	81.08 ± 0.52	5.97 ± 0.67
CPrompt	CVPR'24	87.83 ± 0.13	4.88 ± 0.02	76.70 ± 0.23	6.08 ± 0.19	82.69 ± 0.43	5.30 ± 0.35
Ours	This Paper	**95.52 ± 0.12**	**1.76 ± 0.20**	**79.93 ± 0.19**	3.37 ± 0.41	**88.99 ± 0.15**	**1.46 ± 0.19**

Table 1: Comparison of different continual learning methods on various dataset settings with ImageNet pre-trained ViT. We report results averaged over 3 trials. The best results are marked in red. The second best results are marked in blue.

Moreover, apart from the two-stage inference paradigm, our method can be conducted cyclically multiple times to obtain more accurate aggregation weights, further improving performance.

Overall Optimization

The overall pipeline of our proposed CAPrompt is shown in Fig.2 and the training and inference process is detailed in Algorithm 1. During training, we perform in a two-stage manner as existing methods. Initially, we assign equal task probabilities to different prompts, then get the aggregated prompts ϕ^1 and the new task probability using Eq.2 and Eq.10. The updated aggregated prompt ϕ^2 is then calculated through Eq.2. The loss function with hyperparameters α and β is calculated as:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{con} + \beta \cdot \mathcal{L}_{l}. \tag{11}$$

During inference, as shown in Algorithm 1, we also initialize equal task probabilities for different prompts. The aggregated prompt and task probabilities are then calculated cyclically over a specified number of cycles num. The final aggregated prompt ϕ^{num} is used to prompt the model to get the prediction. Note that existing methods rely on the key-query matching strategy to select the relevant prompt and require two forward passes through the network. Thus, the computation cost for CAPrompt is similar to existing methods when $num = 2$.

Experiments

Experimental Details

Datasets. We follow existing works (Wang et al. 2024a) to evaluate our proposed method on three public datasets, CIFAR-100 (Krizhevsky, Hinton et al. 2009), ImageNet-R (Hendrycks et al. 2021), and CUB200 (Wah et al. 2011). CIFAR-100, ImageNet-R and CUB200 comprise 100, 200 and 200 classes respectively. These datasets are splited into 10 tasks with disjoint classes for incremental learning.

Evaluation Metrics. Following previous works (Wang et al. 2024a), we use final accuracy (ACC) and average forgetting (AF) for evaluation. ACC represents the average accuracy of all the classes that have already been learned. AF estimates the forgetting of previous tasks by calculating the average performance degradation of each classes in different tasks during incremental learning.

Comparison Methods. We compare our methods with various prompt-based pre-trained model based CIL methods, such as L2P (Wang et al. 2022d) , DualPrompt (Wang et al. 2022c), CODA-Prompt (Smith et al. 2023), HiDe-Prompt (Wang et al. 2024a), EvoPrompt (Kurniawan et al. 2024), ConvPrompt (Gao, Cen, and Chang 2024) and Cprompt (Roy et al. 2024). We also include existing methods using other PEFT, such as InfLoRA (Liang and Li 2024), EASE (Zhou et al. 2024a) and SSIAT (Tan et al. 2024).

Implementation Details. For all experiments, we use the ImageNet21K pre-trained ViT-B/16 as the backbone. The parameters are optimized by an Adam optimizer with an initial learning rate of 3e-3 and a batch size of 24. Prefix Tuning Prompts (Wang et al. 2022c), with the prompt length $L_p = 10$ are inserted into all layers for CIFAR-100, CUB200, and into the first nine layers for ImageNet-R with $L_p = 20$. The prompt parameters for each task are initialized using the prompts from the previous task. The weighting parameters of different losses are $\alpha = 5$ and $\beta = 0.2$. The cyclic number num is set to 2 without special explanations for a similar computation cost compared to existing methods. Following HiDe-Prompt, we also maintain one feature per class and replay this feature in subsequent incremental learning tasks to mitigate the forgetting of the classification head. For comparison methods, we reimplement their results according to their released code and the experimental setting reported in their paper. All results are the average performance across 3 runs.

Comparison with SOTA

Main Results. Table.1 represents the results of final accuracy on various datasets. Compared to other prompt-based methods, our CAPrompt achieves the highest accuracy. Specifically, our method achieves final accuracies of **95.52%**, **79.93%**, and **88.99%** on CIFAR-100, ImageNet-R, and CUB200 datasets with the performance gain of

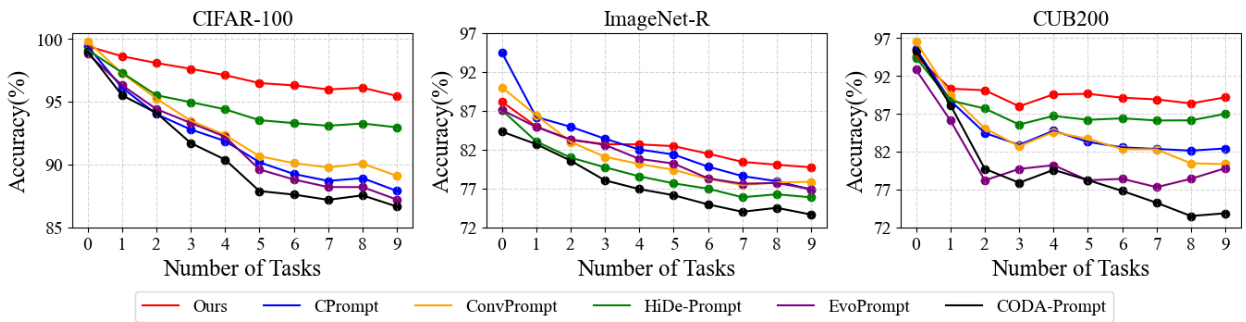

Figure 4: The complete classification accuracy of different methods on each task.

Methods	CIFAR-100		ImageNet-R		CUB200	
	ACC	*AF*	*ACC*	*AF*	*ACC*	*AF*
InfLoRA	86.60	4.87	74.77	6.60	77.45	4.93
EASE	87.58	6.55	76.80	5.07	86.74	2.63
SSIAT	90.71	4.89	79.37	6.96	87.06	7.63
Ours	**95.52**	**1.76**	**79.93**	**3.37**	**88.99**	**1.46**

Table 2: Comparison of LoRA-based and Adapter-based methods on various dataset settings.

Methods	ImageNet-R	
	ACC	*AF*
Base	74.51	3.01
+ Aggregation	78.28	4.41
+ Aggregation + Cyclic	79.16	4.52
+ Aggregation + Cyclic + $\mathcal{L}_{con}$	79.77	3.70
+ Aggregation + Cyclic + $\mathcal{L}_{con}$ + $\mathcal{L}_{l}$	79.93	3.37

Table 3: Ablation study of different components.

2.88%, 1.99%, and 2.15% compared to the second best player. In terms of average forgetting, our results demonstrate the lowest forgetting rate on CIFAR-100 and CUB200. Although HiDe-Prompt exhibits slightly lower forgetting on ImageNet-R, its knowledge acquisition is limited, leading to inferior final accuracy results. Furthermore, we compare our method with the latest LoRA-based and Adapter-based methods in Table.2. Results show that our method also achieves the highest final accuracy and the lowest average forgetting demonstrating its effectiveness.

The outstanding results of our method are attributed to the prompt aggregation strategy, which ensures the consistency of prompts between training and inference. While CODA-Prompt proposed a weighted sum of prompts with an attention-based component-weighting scheme for different prompts, it lacks constraints on the learning of prompt parameters. In contrast, our CAPrompt employs concave and linear constraints, ensuring that the aggregated prompt yields lower prediction errors than using a single task-specific prompt. Additionally, the proposed cyclic prompt weight strategy fully utilizes the knowledge learned by prompts, further improving performance.

Accuracy Curve. To present our results in detail, we present the final accuracy of different methods after different tasks in Fig.4. Notably, with similar accuracy for the initial task, our method achieves the best results across subsequent tasks. This performance gain can be attributed to the robust prompt aggregation strategy and the cyclic prompt weight strategy which effectively leverages the information captured by the prompts.

Ablation Study

Effectiveness of Different Components. The ablation results on ImageNet-R are presented in Table.3. The base model uses task-specific Prefix Tuning Prompts for the first nine layers, with prompt parameters initialized from previous tasks. Our method comprises two strategies and two losses: prompt aggregation, cyclic prompt weight, $\mathcal{L}_{con}$, and $\mathcal{L}_{l}$. Table.3 demonstrates that incorporating prompt aggregation leads to a significant improvement in final accuracy. This is because the prompt aggregation ensures consistency of prompts between the training and inference, thus mitigating knowledge forgetting. Moreover, the aggregated prompt combines similar knowledge from multiple classes and enhances knowledge acquisition. Additionally, the cyclic prompt weight strategy further improves the accuracy by 0.88%, due to its effective utilization of prompts for weight determination. These improvements in knowledge acquisition increase the difficulty of anti-forgetting, thus leading to a slight increase in the metric of average forgetting. Moreover, the concave and linear constraints each contribute to a 0.61% and 0.16% increase in final accuracy and a 0.82% and 0.33% reduction in average forgetting, respectively. This highlights their effectiveness in constraining prompt learning and mitigating knowledge forgetting.

Linear Constraint. To evaluate the impact of $\mathcal{L}_{l}$, we visualize the prompts of layer one for key tokens after training on the ImageNet-R dataset in Fig.6. The number of prompts is $L_p/2 = 10$ for key tokens. The visualization results show that applying $\mathcal{L}_{l}$ effectively aligns the prompts of different tasks in a linear direction and brings them closer together.

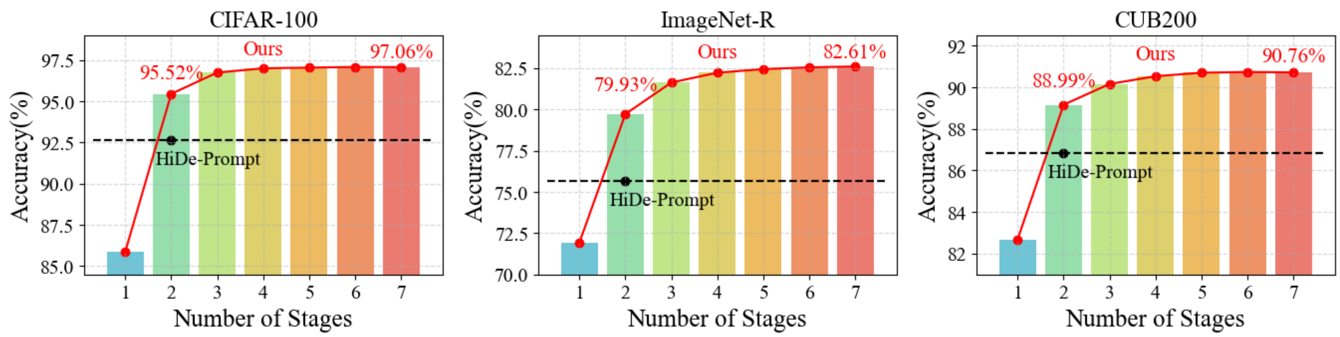

Figure 5: Accuracy of the proposed method increases with the number of cycles (num).

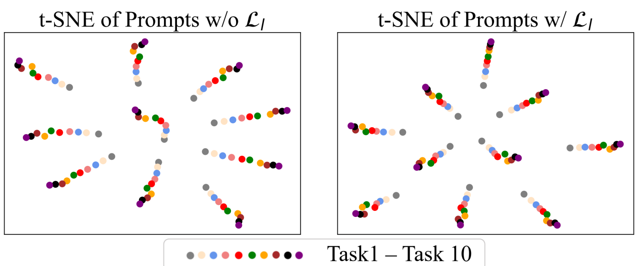

Figure 6: t-SNE visualization of prompts w/o $\mathcal{L}_l$ and w/ $\mathcal{L}_l$.

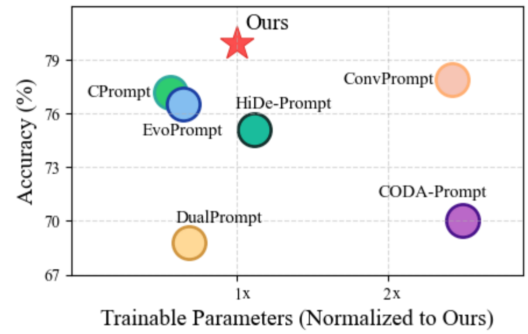

Figure 7: The accuracy and number of trainable parameters for different methods on the ImageNet-R dataset. All trainable parameters are normalized to our method.

This alignment facilitates prompt aggregation by constraining the possible locations of aggregated prompts, which simplifies the training process to meet the concave condition described in Theorem 1. Consequently, this alignment ensures that the aggregated prompt achieves a lower prediction error compared to using a single task-specific prompt.

Cyclic Prompt Weight. As described in the method section, our approach allows for multiple cycles to refine the aggregation weight. For a fair comparison, we set the cyclic number $num = 2$ for the results shown in Table.1, Table.2 and Table.3. In this section, we evaluate the impact of performing additional cycles on prediction accuracy and present the results in Fig.5. Our findings reveal that, with two cycles ($num = 2$), our method achieves superior accuracy compared to the second-best method, HiDe-Prompt. Furthermore, by iterating the cyclic prompt weight adjustment over multiple stages, we observe an additional performance improvement of 1.5%-2%. This enhancement is attributed to the iterative refinement of aggregation weights, which leads to progressively more accurate prompt combinations and consequently better overall performance.

Parameters Comparison. Fig.7 presents the accuracy and the number of trainable parameters for different methods on ImageNet-R datasets, with all trainable parameters normalized according to our method. CODA-Prompt proposed an attention-based component-weighting scheme, and ConvPrompt employs a convolutional prompt generator, both requiring more parameters. Although CPrompt utilizes the fewest parameters, it sequentially prepends prompts to the input tokens, increasing the number of tokens and the computational cost. Our method requires a similar number of trainable parameters compared to other methods but achieves the highest accuracy performance. This superior performance is attributed to our prompt aggregation and cyclic prompt weight strategies, which make full use of the knowledge embedded in the prompts.

Conclusion

In this paper, we present the Cyclic Prompt Aggregation (CAPrompt) method to address the challenges of prompt inconsistency in task ID prediction in Class Incremental Learning (CIL). By aggregating the knowledge of different prompts, CAPrompt eliminates the need for task ID prediction, thereby reducing inconsistencies between training and inference prompts and mitigating knowledge forgetting. The concave constraint and linear constraint are proposed to ensure the aggregated prompt yields lower error than using one task-specific prompt. Additionally, the cyclic weight prediction strategy is proposed to cyclically adjust the prompt weights to more accurate ones, further enhancing performance. Extensive experiments on various datasets demonstrated that CAPrompt significantly outperforms the state-of-the-art methods. Our results underscore the effectiveness of prompt aggregation and the importance of cyclically utilizing the learned prompts in CIL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376011).

References

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the European Conference on Computer Vision*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Gao, Z.; Cen, J.; and Chang, X. 2024. Consistent Prompting for Rehearsal-Free Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28463–28473.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hu, Z.; Li, Y.; Lyu, J.; Gao, D.; and Vasconcelos, N. 2023. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jung, D.; Han, D.; Bang, J.; and Song, H. 2023. Generating instance-level prompts for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11847–11857.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kurniawan, M. R.; Song, X.; Ma, Z.; He, Y.; Gong, Y.; Qi, Y.; and Wei, X. 2024. Evolving Parameterized Prompt Memory for Continual Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13301–13309.

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

Li, Q.; Peng, Y.; and Zhou, J. 2024a. FCS: Feature Calibration and Separation for Non-Exemplar Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28495–28504.

Li, Q.; Peng, Y.; and Zhou, J. 2024b. Progressive Prototype Evolving for Dual-Forgetting Mitigation in Non-Exemplar Online Continual Learning. In *ACM Multimedia 2024*.

Li, Q.; Xu, K.; Peng, Y.; and Zhou, J. 2024. Exemplar-Free Lifelong Person Re-identification via Prompt-Guided Adaptive Knowledge Consolidation. *International Journal of Computer Vision*, 1–16.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40.

Liang, Y.-S.; and Li, W.-J. 2024. InfLoRA: Interference-Free Low-Rank Adaptation for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23638–23647.

Liu, Y.; Schiele, B.; and Sun, Q. 2021. Rmm: Reinforced memory management for class-incremental learning. *Advances in Neural Information Processing Systems*, 34.

Liu, Z.; Peng, Y.; and Zhou, J. 2024. Compositional Prompting for Anti-Forgetting in Domain Incremental Learning. *International Journal of Computer Vision*, 1–18.

Luo, Z.; Liu, Y.; Schiele, B.; and Sun, Q. 2023. Class-incremental exemplar compression for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Prabhu, A.; Torr, P. H.; and Dokania, P. K. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *Proceedings of the European conference on computer vision*. Springer.

Qiao, J.; Tan, X.; Chen, C.; Qu, Y.; Peng, Y.; Xie, Y.; et al. 2023. Prompt Gradient Projection for Continual Learning. In *The Twelfth International Conference on Learning Representations*.

Roy, A.; Moulick, R.; Verma, V. K.; Ghosh, S.; and Das, A. 2024. Convolutional Prompting meets Language Models for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23616–23626.

Smith, J.; Hsu, Y.-C.; Balloch, J.; Shen, Y.; Jin, H.; and Kira, Z. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11909–11919.

Tan, Y.; Zhou, Q.; Xiang, X.; Wang, K.; Wu, Y.; and Li, Y. 2024. Semantically-Shifted Incremental Adapter-Tuning is A Continual ViTransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23252–23262.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, F.-Y.; Zhou, D.-W.; Liu, L.; Ye, H.-J.; Bian, Y.; Zhan, D.-C.; and Zhao, P. 2022a. BEEF: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The Eleventh International Conference on Learning Representations*.

Wang, F.-Y.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2022b. Foster: Feature boosting and compression for class-incremental learning. In *Proceedings of the European conference on computer vision*.

Wang, L.; Xie, J.; Zhang, X.; Huang, M.; Su, H.; and Zhu, J. 2024a. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024b. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Y.; Ma, Z.; Huang, Z.; Wang, Y.; Su, Z.; and Hong, X. 2023. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10209–10217.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022c. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 631–648. Springer.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022d. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Xu, K.; Zou, X.; Peng, Y.; and Zhou, J. 2024. Distribution-aware Knowledge Prototyping for Non-exemplar Lifelong Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16604–16613.

Xu, K.; Zou, X.; and Zhou, J. 2024. LSTKC: Long Short-Term Knowledge Consolidation for Lifelong Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16202–16210.

Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; and He, Y. 2024. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23219–23230.

Zhou, D.-W.; Sun, H.-L.; Ye, H.-J.; and Zhan, D.-C. 2024a. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23554–23564.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024b. Class-Incremental Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, D.-W.; Wang, Q.-W.; Ye, H.-J.; and Zhan, D.-C. 2022. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*.