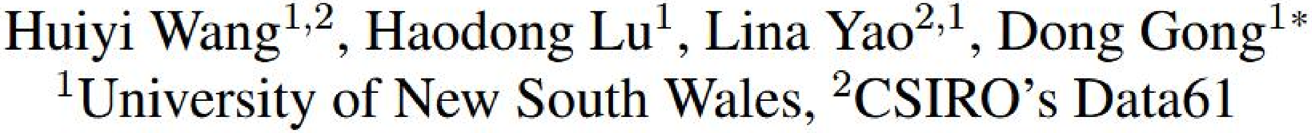
Self-Expansion of Pre-trained Models with Mixture of Adapters for Continual Learning

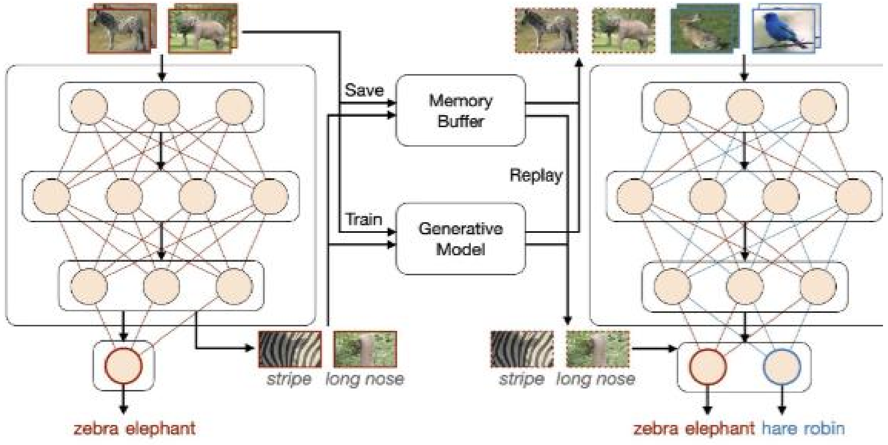


Huiyi Wang[1,2], Haodong Lu[1], Lina Yao[2,1], Dong Gong[1]*
[1]University of New South Wales, [2]CSIRO's Data61

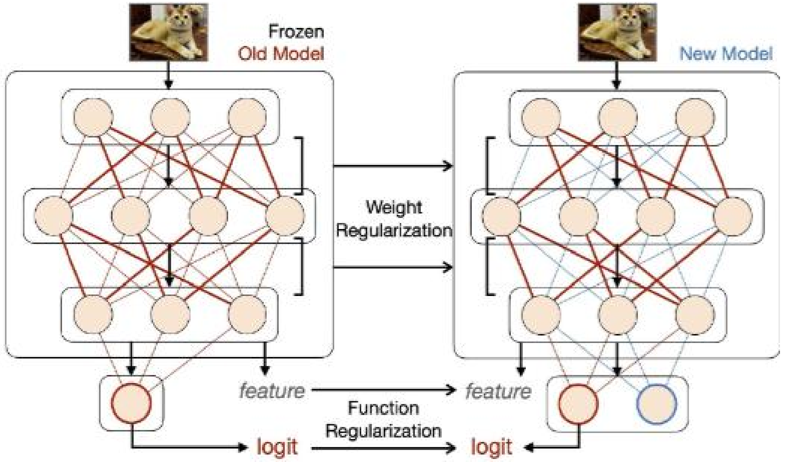
CVPR 2025

Background

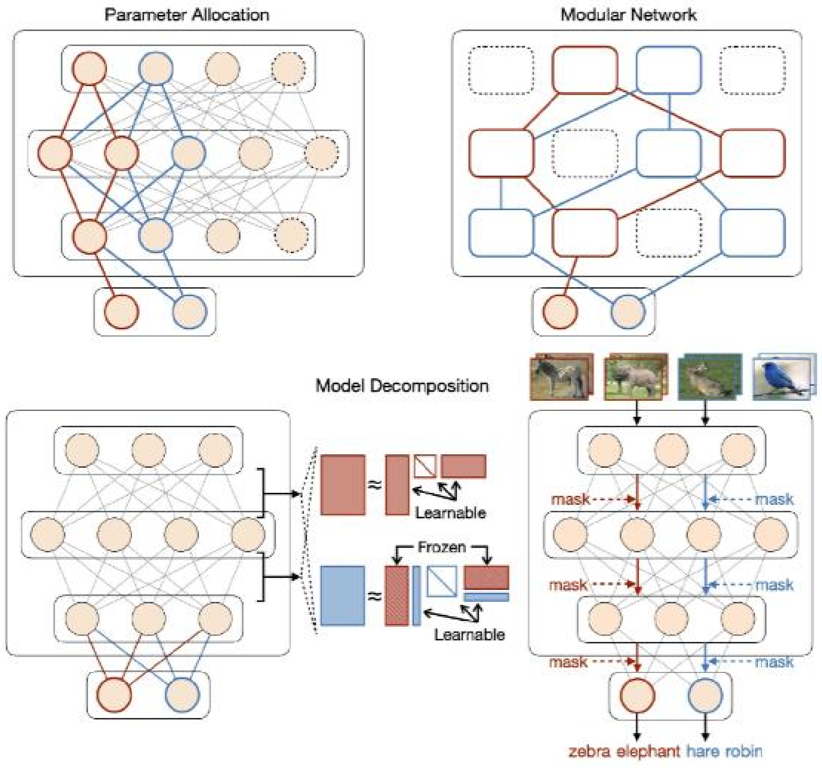
- Traditional continual learning methods



Replay-based methods



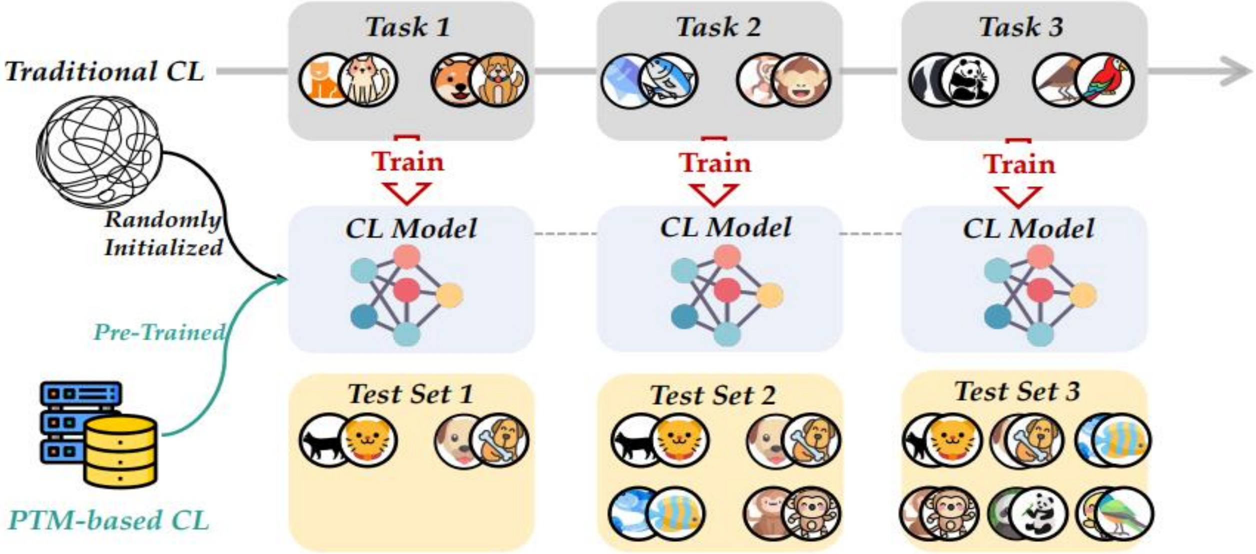
Regularization-based methods



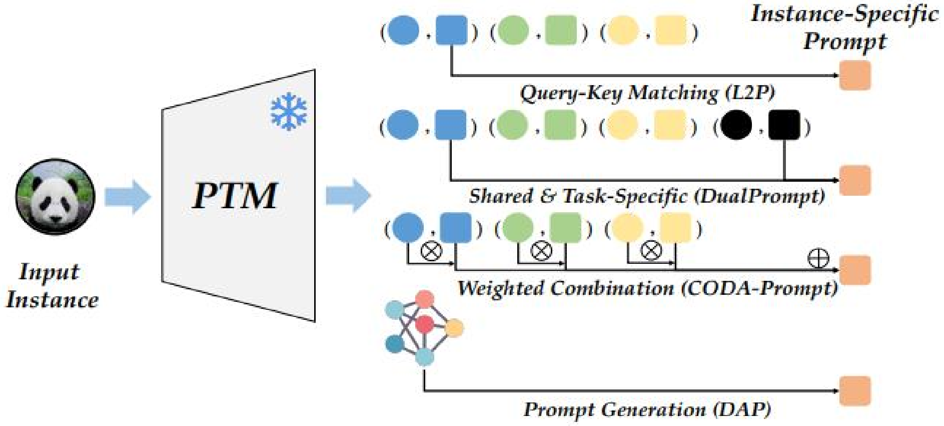
Architecture-based methods

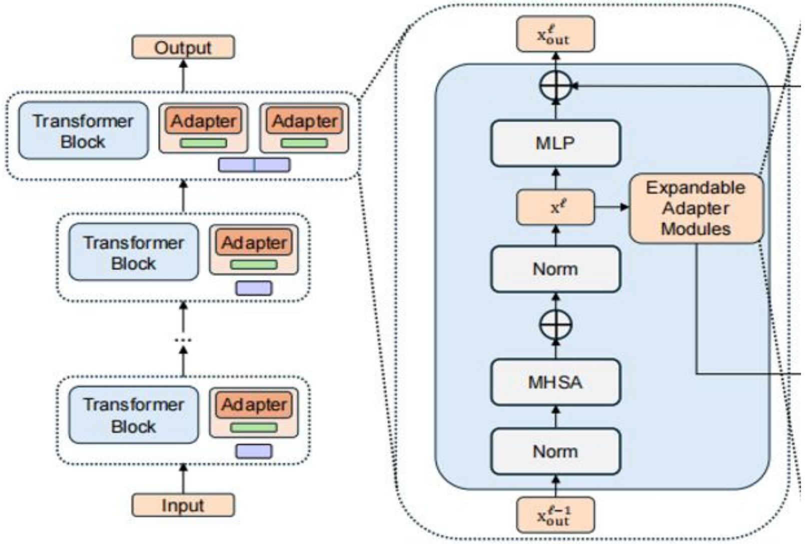
1. 基于**重放**的方法受限于内存开销和隐私问题。
2. 基于**正则化**的方法通常会限制模型学习新任务的能力。
3. 基于**架构**的方法可以有效减少遗忘，但通常会导致模型快速增长和结构复杂性增加。

Background



大多数现有的基于 PTM 的 CL 方法保持 PTM 固定不变，以维持稳定的表示并减轻灾难性遗忘，并通过参数高效的微调，利用新扩展的prompt或apapter进行持续适应。





Background

现有基于 PTM 的 Adapter / Prompt 持续学习方法，主流上可以归纳为两类：
（1）固定一组 Adapter/Prompt 共享；
（2）每个任务新增一组 Adapter/Prompt（任务级扩展）。
固定模块微调优点：避免遗忘；缺点：适应新任务能力弱→CL能力受限
每任务增加新模块优点：新任务能力强；缺点：模型不断膨胀，不同任务间知识难以共享 → 低效

➢ 现有基于 PTM 的持续学习方法的主要不足:

1. 固定 Prompt / Adapter → 适应能力受限;
2. 任务级扩展 → 参数线性增长 + 知识难共享;
3. 稳定性–可塑性平衡不可控。

Motivation

➢ 作者针对这些不足的创新方法SEMA:

1. 按需自扩展；
2. 模块化 Adapter + 知识复用；
3. 可扩展的加权路由器。

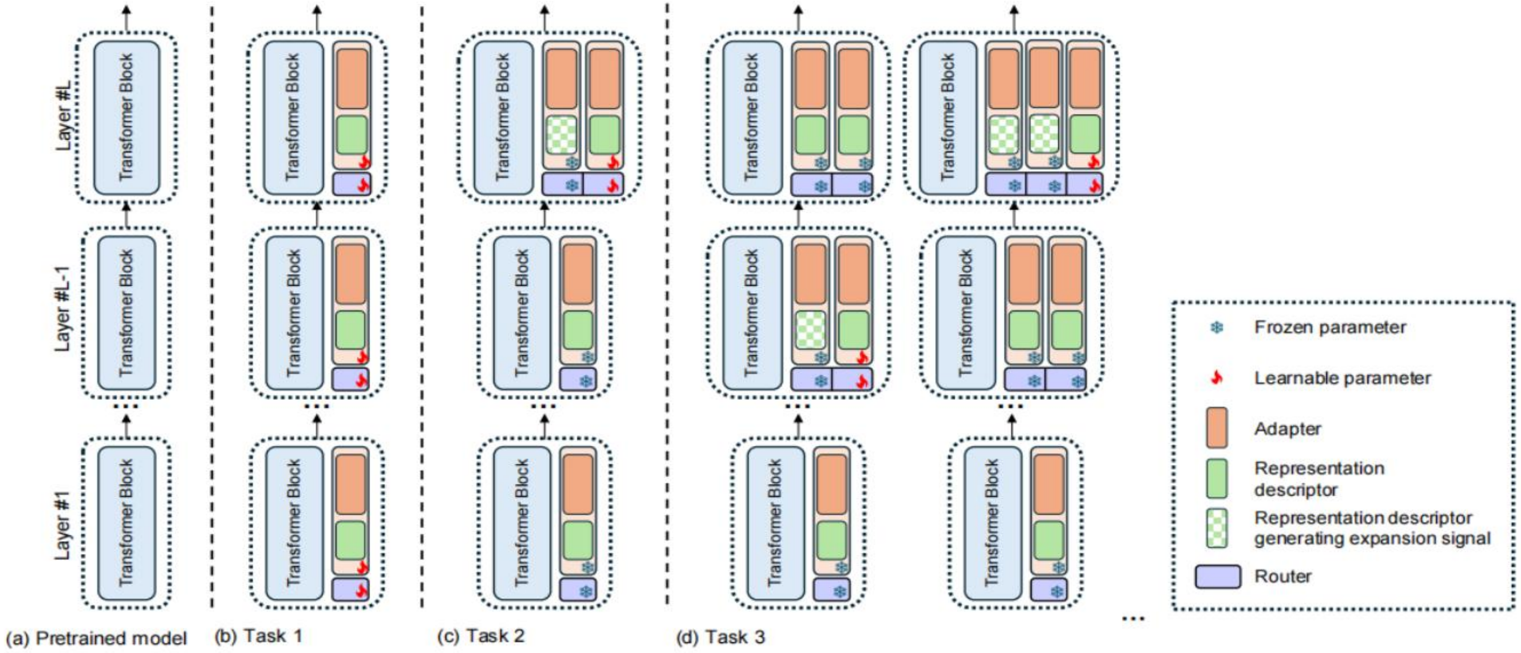


Figure 1. An example of the self-expansion process. (a) The PTM (*i.e.*, ViT) with L transformer layers at the initial point of CL. (b) The first session adaptation – at Task 1, a modular adapter and a (dummy) router is added and trained in each transformer layer. (c) The modular adapters and routers added in the previous step (Task 1) are frozen to alleviate forgetting. When Task 2 arrives, *only* the representation descriptor in the L-th layer detects feature distribution shift (with novel patterns) and generates *expansion signal*. A new module is added and trained in the L-th layer, with the router expanded and updated. (d) At Task 3, new adapter is added at $L-1$-th layer after the expansion signal is firstly generated. In this demo example, the expansion is triggered and produced again in the L-th layer, following the expansion in the $L-1$-th layer. If a task does not trigger expansion signal in any layer (implying no significantly different pattern), expansion would not happen, and existing adapters would be reused. More discussions are in Appendix A.1.

Methods

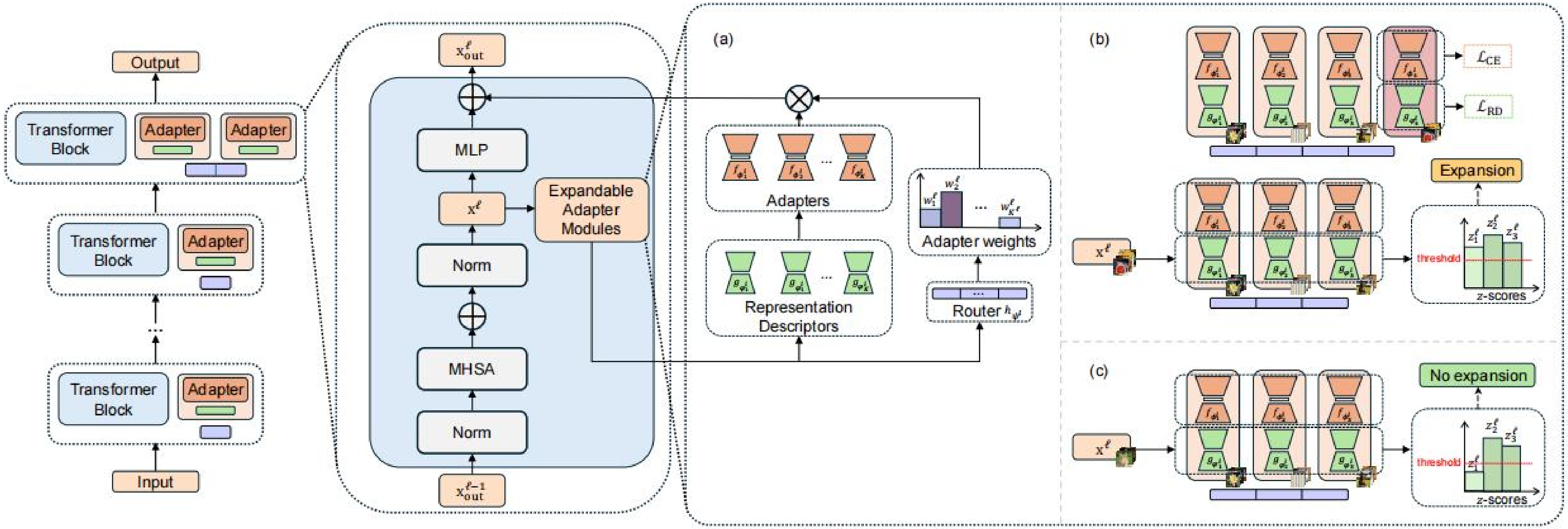


Figure 2. Overview of the model architecture. (a) shows the structure of expandable adapter modules with adapters, RDs and router. (b) shows the scenario where expansion is triggered by representations with distribution different to previous tasks, estimated by RD. RDs are trained to align with the feature distribution of the corresponding task via only $\mathcal{L}_{\mathrm{RD}}$, unaffected by gradients from the classification loss. (c) shows the scenario where incoming distribution can be handled by previously added modules, resulting in no expansion and adapter reuse.

Methods

1、Representation-Aware Modular Adapter

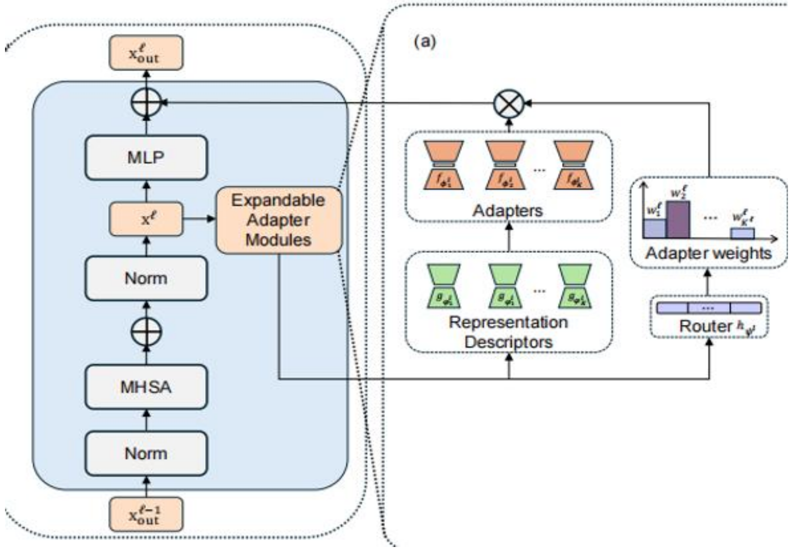
（1）Functional adapter

$$f_{\phi_k^l}(\mathbf{x}^l) = \mathrm{ReLU}(\mathbf{x}^l \cdot \mathbf{W}^l_{\mathrm{down},k}) \cdot \mathbf{W}^l_{\mathrm{up},k}, \qquad (1)$$

$x \in \mathbb{R}^d$

$W_{\mathrm{down}} \in \mathbb{R}^{d \times r}$

$W_{\mathrm{up}} \in \mathbb{R}^{r \times d}$



Methods

（2） Representation descriptor

$$\mathcal{L}^{l}_{\mathrm{RD},k}(x) = \sum\nolimits_{\mathbf{x}\in\mathcal{X}^{l}_{k}} \|\mathbf{x} - g_{\varphi^{l}_{k}}(\mathbf{x})\|^{2}_{2}. \qquad (2)$$

$$r^{l}_{k} = \|x_i - g^{l}_{k}(x_i)\|^{2}_{2}$$

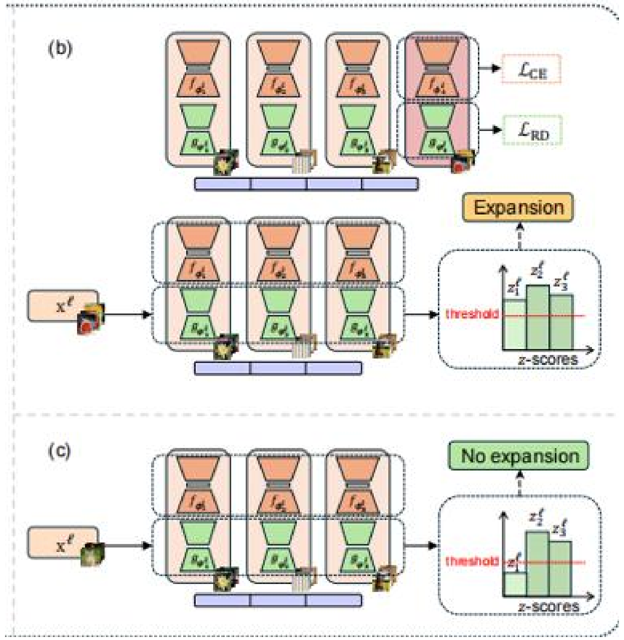
$$\hat{x} = g(x) = \mathrm{Dec}(\mathrm{Enc}(x))$$

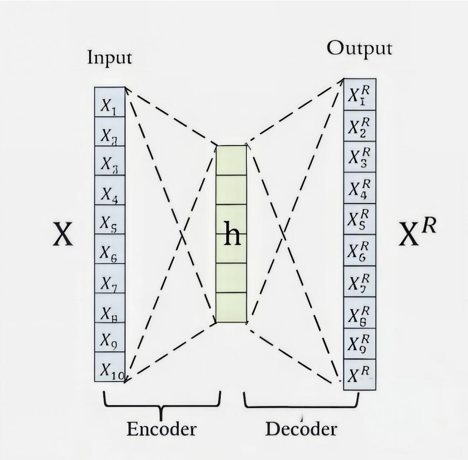
x：该层 MLP 的输入特征

$g^{l}_{k}(x)$：AE 的重构输出

$$z^{l}_{k} = \frac{r^{l}_{k} - \mu^{l}_{k}}{\sigma^{l}_{k}}$$

r^{l}_{k}	当前输入 x_l 经过第 k 个 RD 重建的 **重建误差**
μ^{l}_{k}	训练时该 RD 的 **平均重建误差**
σ^{l}_{k}	训练时该 RD 的 **重建误差标准差**





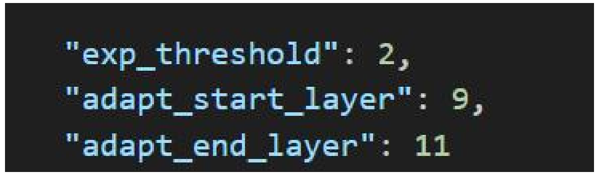
AntoEncoder结构示意图

Methods

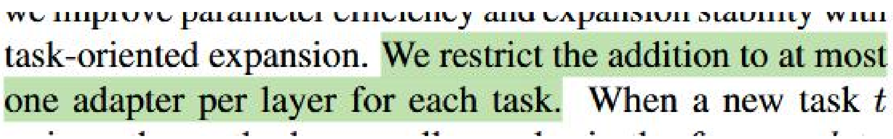
（2） Representation descriptor

$$\mathcal{L}^{l}_{\mathrm{RD},k}(x) = \sum\nolimits_{\mathbf{x}\in\mathcal{X}^{l}_{k}} \|\mathbf{x} - g_{\varphi^{l}_{k}}(\mathbf{x})\|^{2}_{2}. \qquad (2)$$

这个损失的意义是让这个 AE 只对“当前这个任务的特征分布”有很好的重构能力。



```
"exp_threshold": 2,
"adapt_start_layer": 9,
"adapt_end_layer": 11
```



we improve parameter efficiency and expansion stability with task-oriented expansion. We restrict the addition to at most one adapter per layer for each task. When a new task t

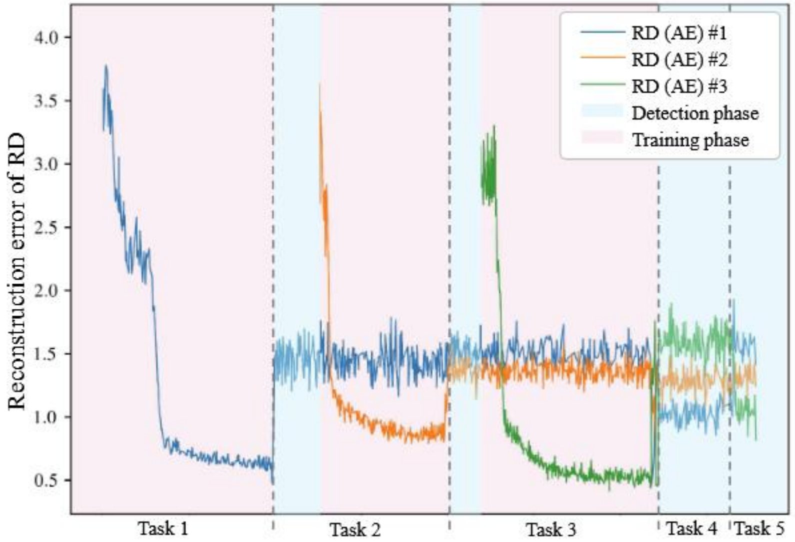
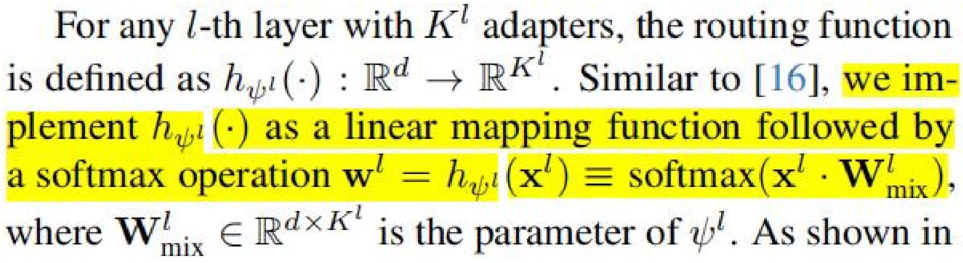


Figure 4. Reconstruction error during training to show the dynamic expansion process. Expansion occurs for Tasks 1, 2, and 3, while no expansion is triggered for Tasks 4 and 5 due to no detected distribution shift.

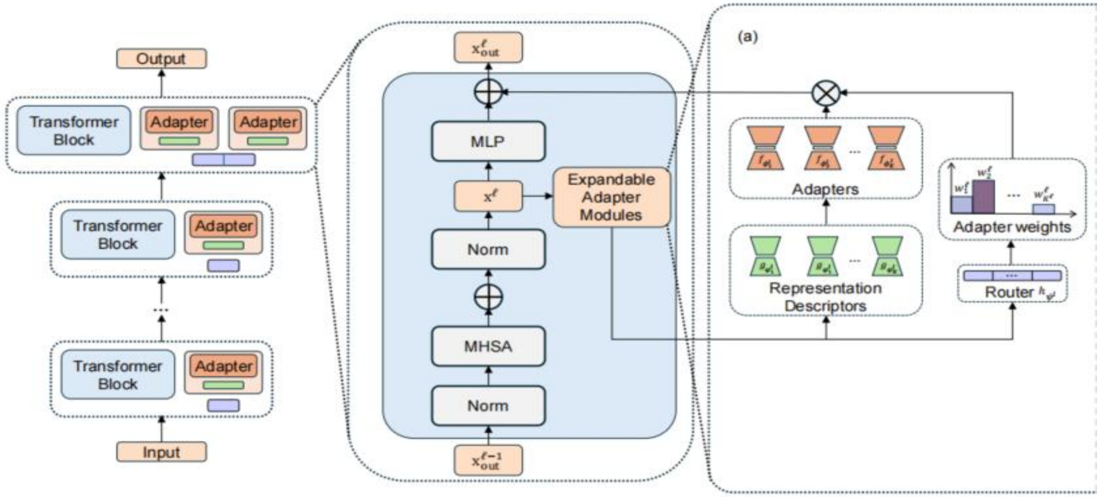
Methods

2、Expandable Weighting Router for Mixture Usage of Adapters



For any l-th layer with K^l adapters, the routing function is defined as $h_{\psi^l}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{K^l}$. Similar to [16], we implement $h_{\psi^l}(\cdot)$ as a linear mapping function followed by a softmax operation $\mathbf{w}^l = h_{\psi^l}(\mathbf{x}^l) \equiv \text{softmax}(\mathbf{x}^l \cdot \mathbf{W}^l_{\text{mix}})$, where $\mathbf{W}^l_{\text{mix}} \in \mathbb{R}^{d \times K^l}$ is the parameter of ψ^l. As shown in

$$\mathbf{x}^l_{\text{out}} = \text{MLP}(\mathbf{x}^l) + \sum\nolimits_{k=1}^{K^l} w^l_k \cdot f_{\phi^l_k}(\mathbf{x}^l). \qquad (3)$$



3、Continual Learning Objective of SEMA

$$\min_{\{\phi^l_k\},\{\psi^l\},\{\varphi^l_k\}} \sum\nolimits_{t=1}^{T} \mathbb{E}_{(x,y)\in D^t} \Big[\mathcal{L}_{\text{CE}}(F_{\{\phi^l_k\},\{\psi^l\}}(x),\ y) + \sum\nolimits_{l=1}^{L} \sum\nolimits_{k=1}^{K^l} \mathcal{L}^l_{\text{RD},k}(x;\varphi^l_k) \Big].$$

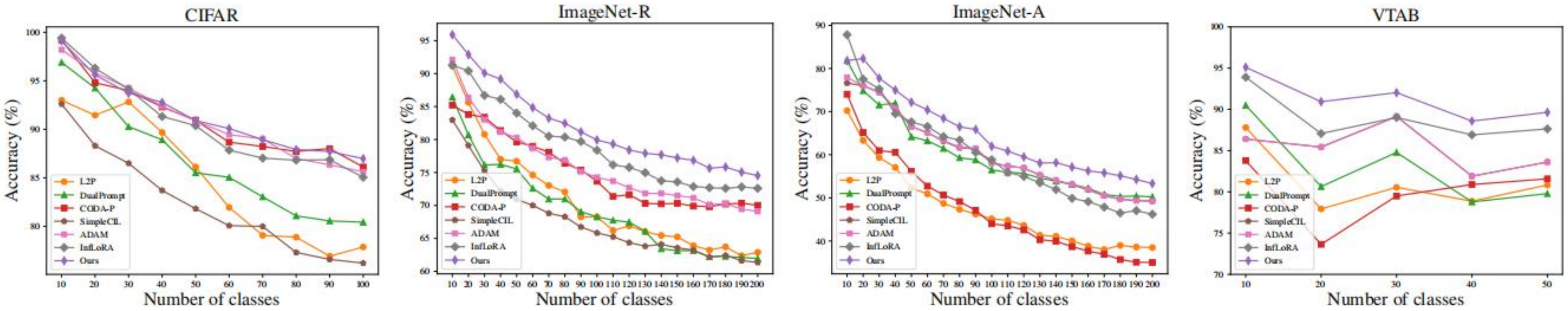
Experiments

Method	CIFAR-100		5-Task IN-R		10-Task IN-R		20-Task IN-R		ImageNet-A		VTAB	
	$\bar{\mathcal{A}}$	$\mathcal{A}_N$	$\bar{\mathcal{A}}$	$\mathcal{A}_N$	$\bar{\mathcal{A}}$	$\mathcal{A}_N$	$\bar{\mathcal{A}}$	$\mathcal{A}_N$	$\bar{\mathcal{A}}$	$\mathcal{A}_N$	$\bar{\mathcal{A}}$	$\mathcal{A}_N$
FT Adapter	47.88	30.9	53.91	41.23	45.31	30.93	38.51	24.22	29.78	17.64	59.98	43.50
L2P	84.77	77.87	77.40	73.59	66.97	62.72	70.67	62.90	47.16	38.48	81.19	80.83
DualPrompt	86.60	80.43	76.39	72.29	72.83	66.75	62.33	61.97	59.54	50.23	82.89	79.79
CODA-P	**91.55**	86.11	81.63	76.98	81.11	75.25	75.00	70.02	47.29	35.02	79.88	81.58
SimpleCIL	82.31	76.21	65.83	61.31	67.09	61.35	67.59	61.35	60.05	49.24	85.29	83.61
ADAM	90.55	85.62	79.91	74.25	79.11	73.15	75.84	69.10	60.15	49.24	85.29	83.61
InfLoRA	90.51	85.05	78.58	72.58	81.39	75.32	78.87	72.60	59.71	46.21	88.90	87.63
SEMA	91.37	**86.98**	**84.75**	**79.78**	**83.56**	**78.00**	**81.75**	**74.53**	**64.53**	**53.32**	**91.26**	**89.64**

Table 1. Comparison with ViT-based CL methods in CIL. All models adopt ViT-B/16-IN1K as the backbone.

$$\mathcal{A}_N = \frac{1}{N}\sum_{i=1}^{N}\mathcal{A}_{i,N},$$

$$\bar{\mathcal{A}} = \frac{1}{N}\sum_{t=1}^{N}\mathcal{A}_t.$$



CIFAR

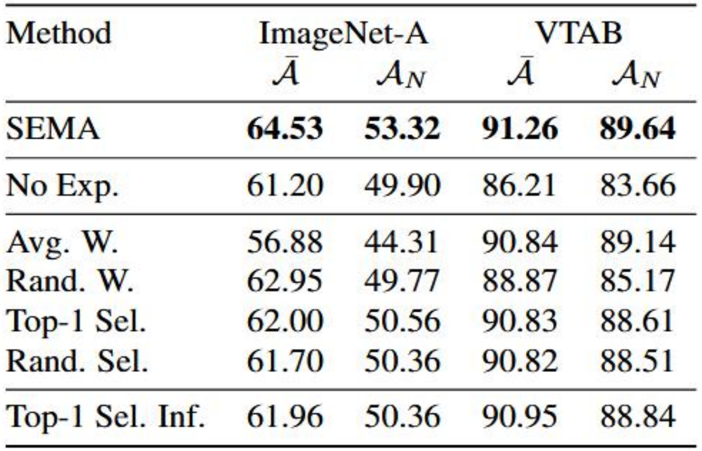
ImageNet-R

ImageNet-A

VTAB

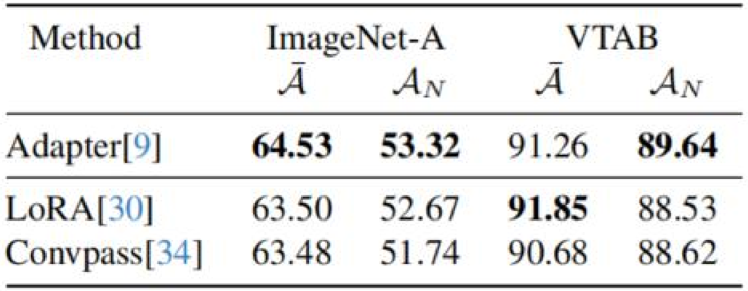
Figure 3. Incremental performance of different methods on class-incremental learning benchmarks.

Experiments



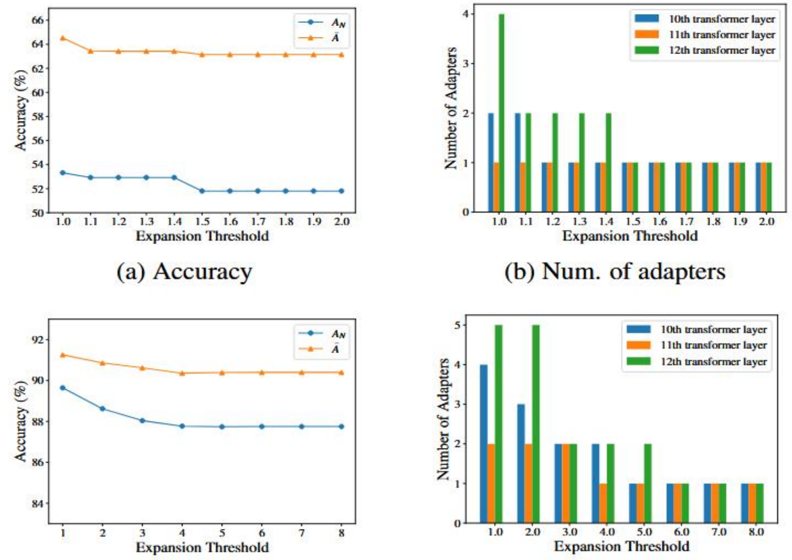
Method	ImageNet-A		VTAB	
	$\bar{A}$	$\mathcal{A}_N$	$\bar{A}$	$\mathcal{A}_N$
SEMA	**64.53**	**53.32**	**91.26**	**89.64**
No Exp.	61.20	49.90	86.21	83.66
Avg. W.	56.88	44.31	90.84	89.14
Rand. W.	62.95	49.77	88.87	85.17
Top-1 Sel.	62.00	50.56	90.83	88.61
Rand. Sel.	61.70	50.36	90.82	88.51
Top-1 Sel. Inf.	61.96	50.36	90.95	88.84

Table 2. Ablation studies on adapter expansion and composing.



Method	ImageNet-A		VTAB	
	$\bar{A}$	$\mathcal{A}_N$	$\bar{A}$	$\mathcal{A}_N$
Adapter[9]	**64.53**	**53.32**	91.26	**89.64**
LoRA[30]	63.50	52.67	**91.85**	88.53
Convpass[34]	63.48	51.74	90.68	88.62

Table 3. Different adapter variants.



(a) Accuracy

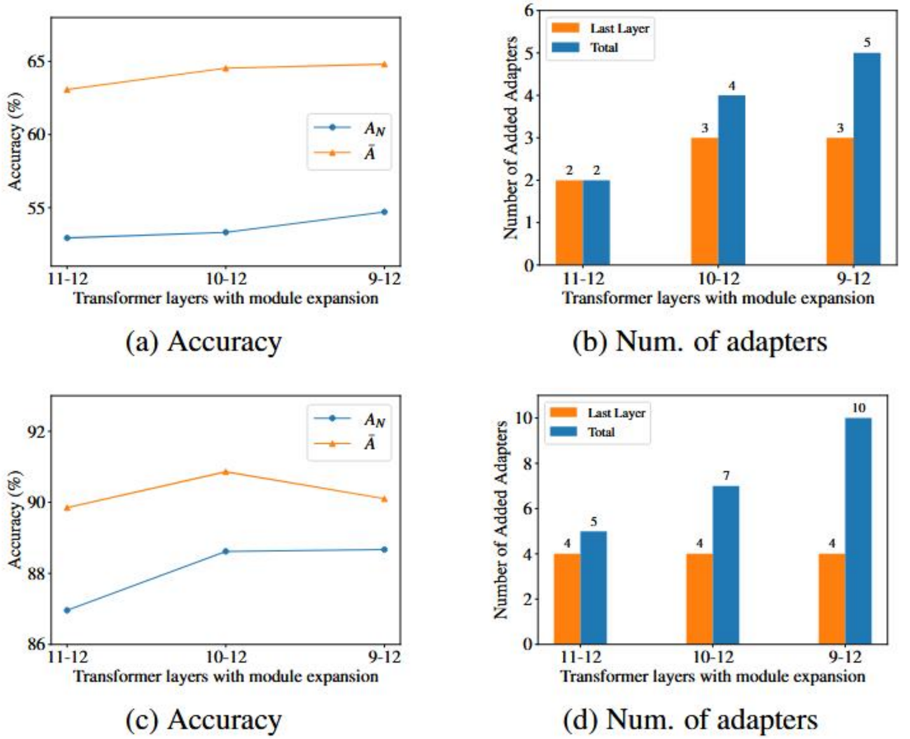
(b) Num. of adapters

(c) Accuracy

(d) Num. of adapters

Figure 6. Analysis of the impact of expansion threshold with (a)(b) ImageNet-A and (c)(d) VTAB. (a) and (c) show that SEMA can produce good accuracy stably with slight variation w.r.t. varying expansion threshold. (b) and (d) report how the number of added adapters (on the specific Transformer layers #10, #11, #12) changes with the varying threshold values, corresponding to (a) and (c), respectively. The proposed method is insensitive to the threshold. Adding more adapters may lead to higher accuracy, a proper threshold can achieve a balance between performance and model size.

Experiments



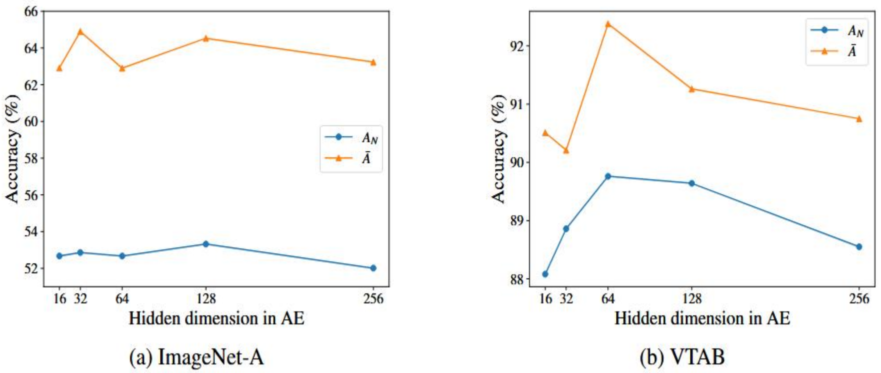
(a) Accuracy

(b) Num. of adapters

(c) Accuracy

(d) Num. of adapters

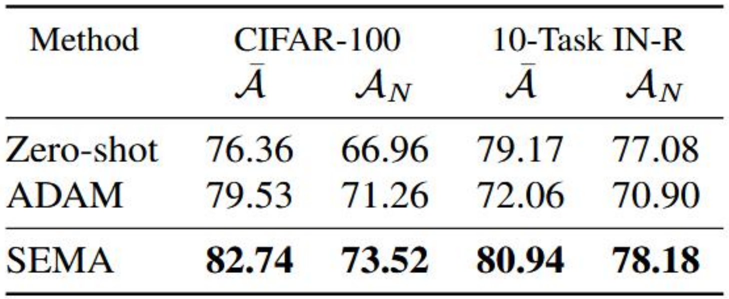
Figure 7. Analysis of the effect of multi-layer expansion, with (a)(b) ImageNet-A and (c)(d) VTAB. By enabling automatic self-expansion on multiple transformer layers, SEMA can achieve better performance than restricting that on a single layer.



(a) ImageNet-A

(b) VTAB

Figure 13. Ablation on representation descriptor.



Method	CIFAR-100		10-Task IN-R	
	$\bar{A}$	$\mathcal{A}_N$	$\bar{A}$	$\mathcal{A}_N$
Zero-shot	76.36	66.96	79.17	77.08
ADAM	79.53	71.26	72.06	70.90
SEMA	**82.74**	**73.52**	**80.94**	**78.18**

Table 15. Performance on pre-trained CLIP model.

Thanks