基于模型兼容的增量学习方法探究

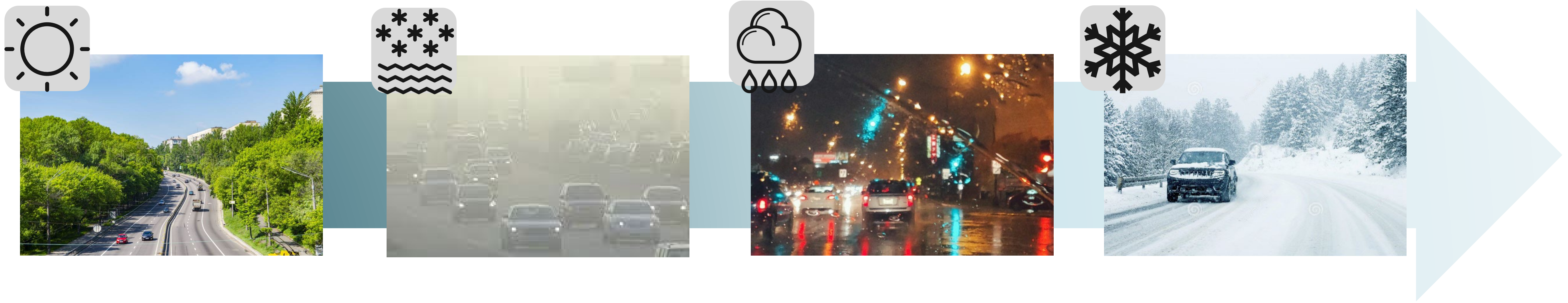
周大蔚
南京大学
zhoudw@nju.edu.cn

持续学习/增量学习

......

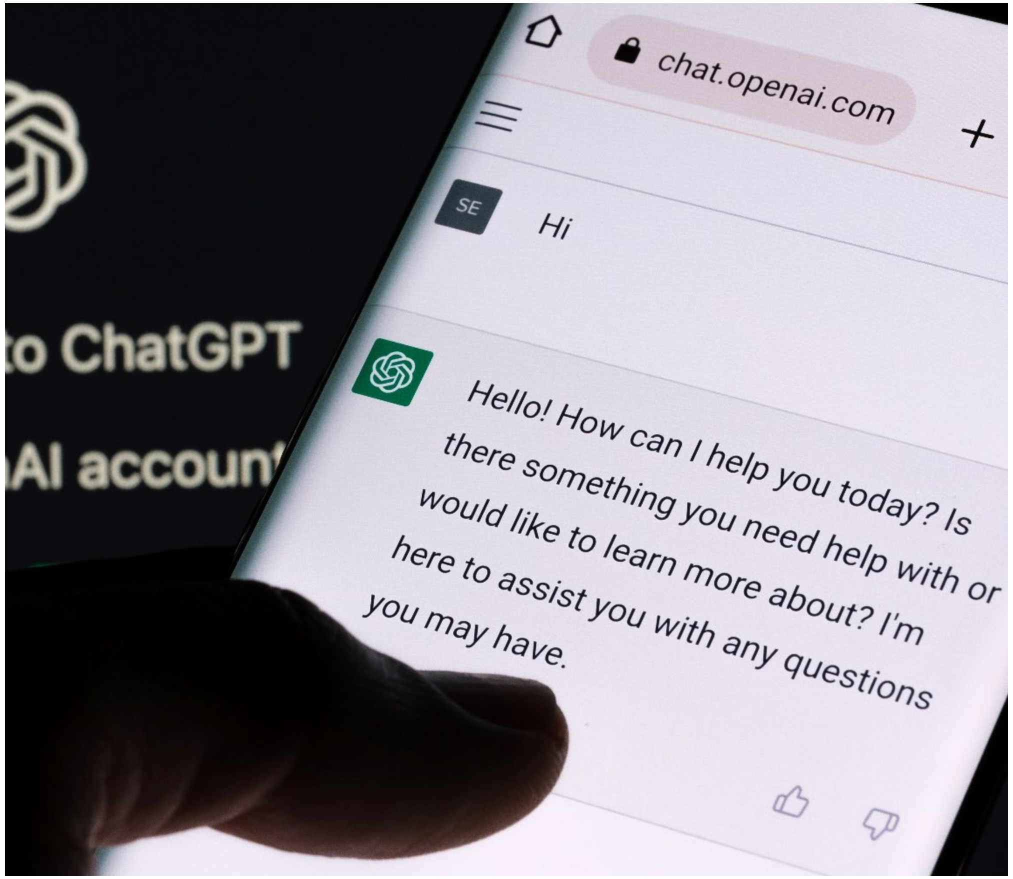
封闭世界训练

开放世界应用



新的场景不断出现，自动驾驶系统需要更新

持续学习/增量学习



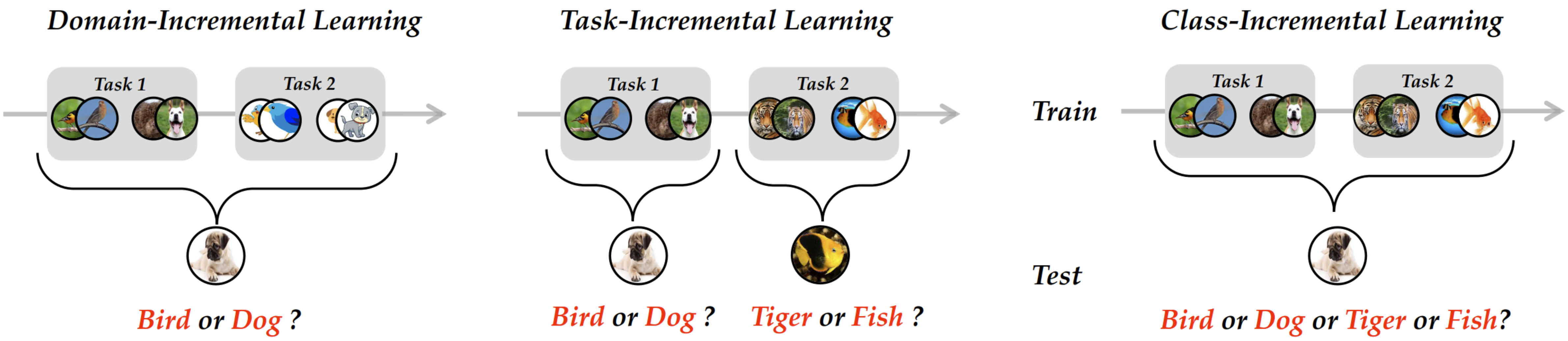
CC How long ago was your last update?

As of my knowledge cut-off date in September 2021, I'm unable to access real-time updates or information beyond that time. Therefore, I may not have the most current information or recent updates. Feel free to ask about information available up until that date!

新闻不断产生，GPT模型需要不断更新

持续学习/增量学习的类型

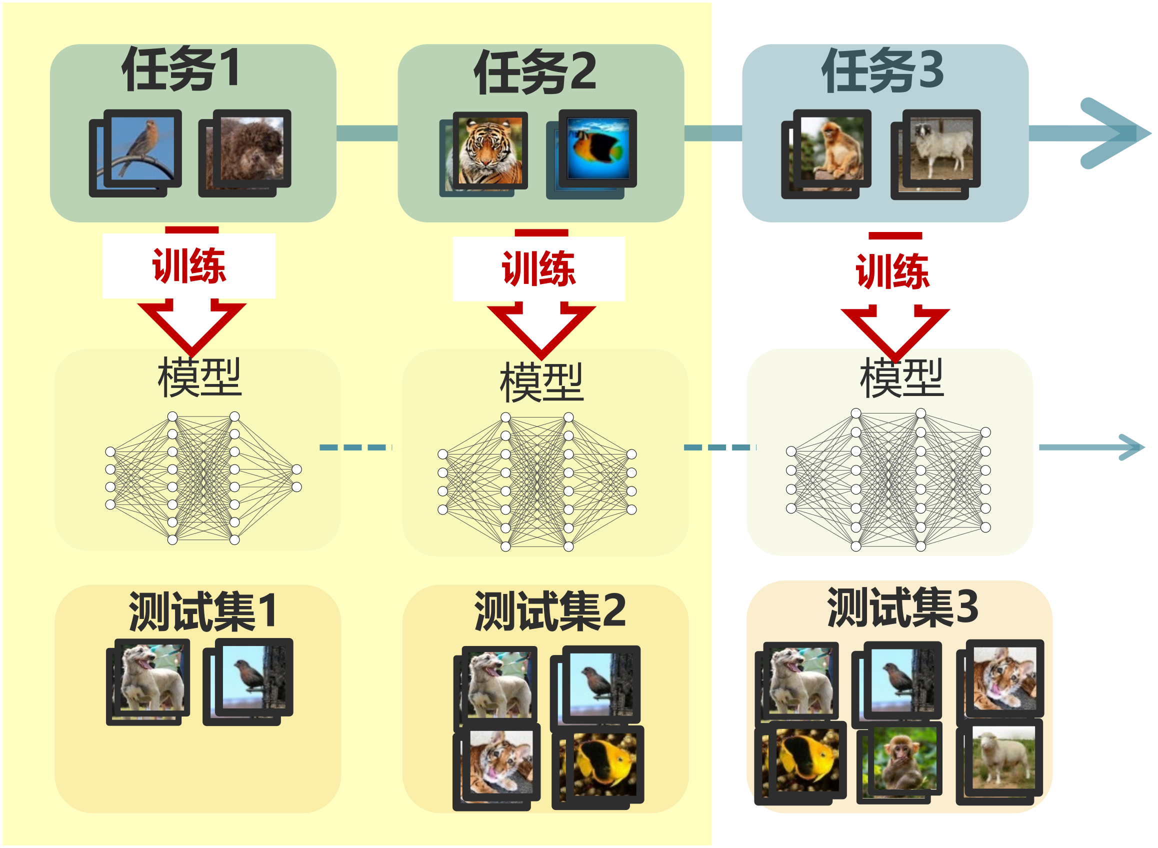
- 增量学习（Incremental Learning）：应对新数据，模型持续扩展

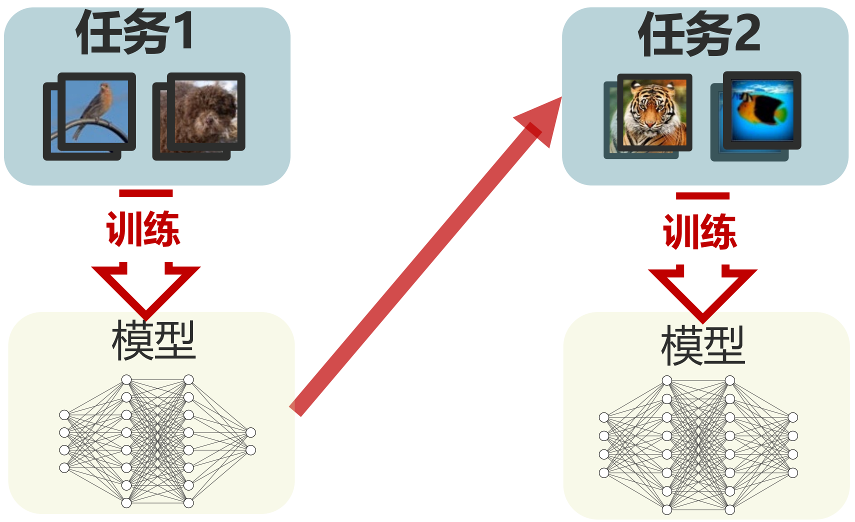


Zhou et al., *Class-Incremental Learning: A Survey*. TPAMI 2024

类别增量学习

- 类别增量学习（Class-Incremental Learning）：模型扩展能够识别新类

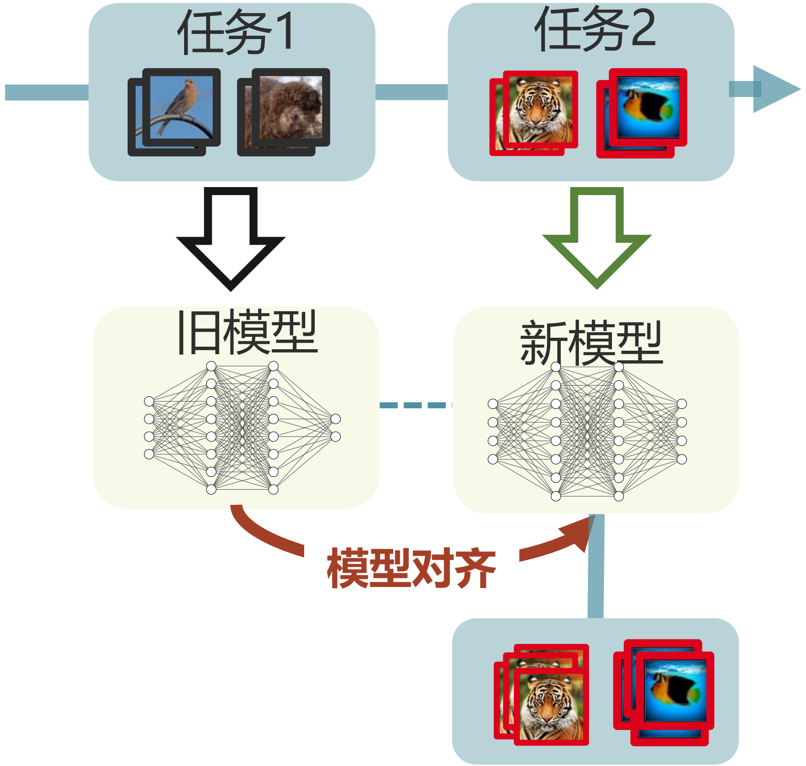




已有的模型在新类数据上微调？

如何不遗忘地扩充模型的能力?

- 增量学习中，需要同时满足
  - 扩充模型对新类别的识别能力
  - 降低模型对已有类别的灾难性遗忘（catastrophic forgetting）



➢ 参数正则化 [Kirkpatrick et al. PNAS'17] [Friedemann et al. ICML'17]

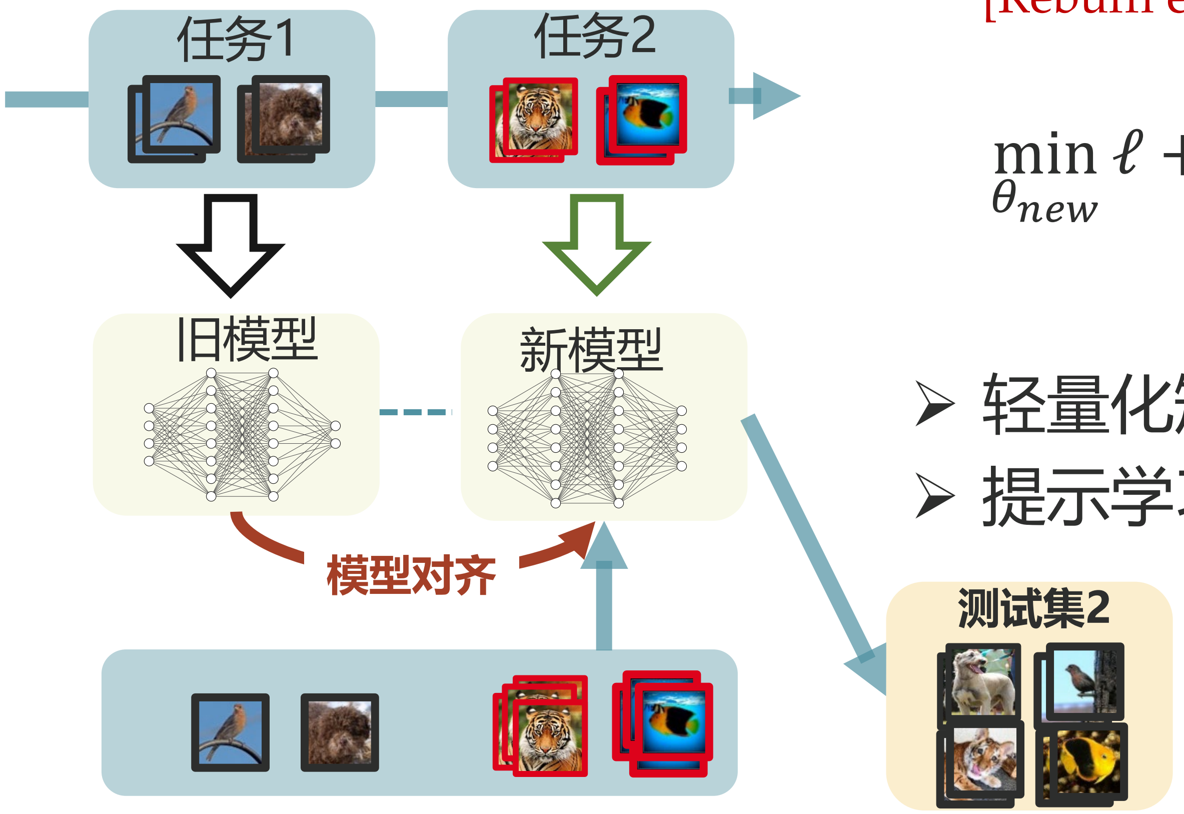
$$\min_{\theta_{new}} \ell - SIM(\boldsymbol{\theta}_{old}, \boldsymbol{\theta}_{new})$$

➢ 知识蒸馏 [Li et al. TPAMI'17] [Baek et al. NeurIPS'22]

$$\min_{\theta_{new}} \ell - \sum_{\boldsymbol{x}_{new}} SIM\left(f_{\boldsymbol{\theta}_{old}}(\boldsymbol{x}_{new}), f_{\boldsymbol{\theta}_{new}}(\boldsymbol{x}_{new})\right)$$

模型难以**平衡**旧类和新类的预测

如何不遗忘地扩充模型的能力?



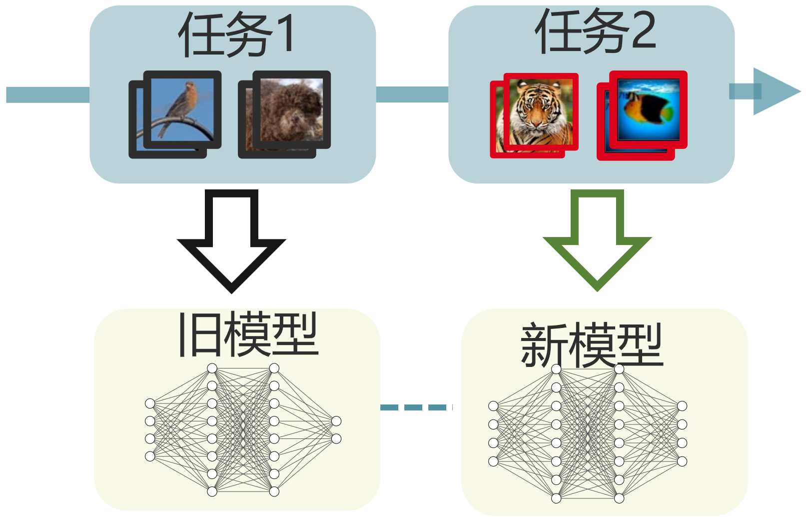
范例集（exemplar set）

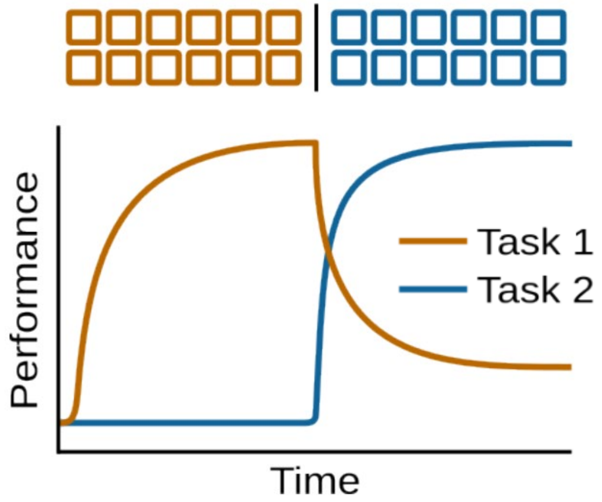
- 存储少量旧类样本， “复刻”已有模型的能力
  [Rebuffi et al. CVPR’17] [Pietro et al. NeurIPS’20]

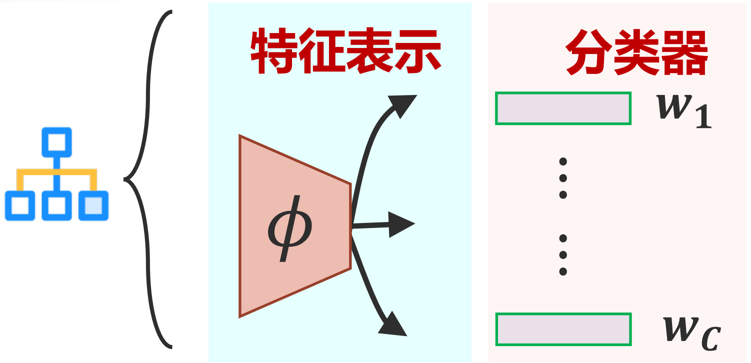
$$\min_{\theta_{new}} \ell + \sum_{\boldsymbol{x} \in \boldsymbol{x}_{new} \cup \boldsymbol{x}_{old}} -SIM\left(f_{\boldsymbol{\theta}_{old}}(\boldsymbol{x}), f_{\boldsymbol{\theta}_{new}}(\boldsymbol{x})\right)$$

- 轻量化矫正 [Wu et al. CVPR’19] [Pham et al. ICLR’22]
- 提示学习 [Wang et al. CVPR’22] [Wang et al. NeurIPS’22]

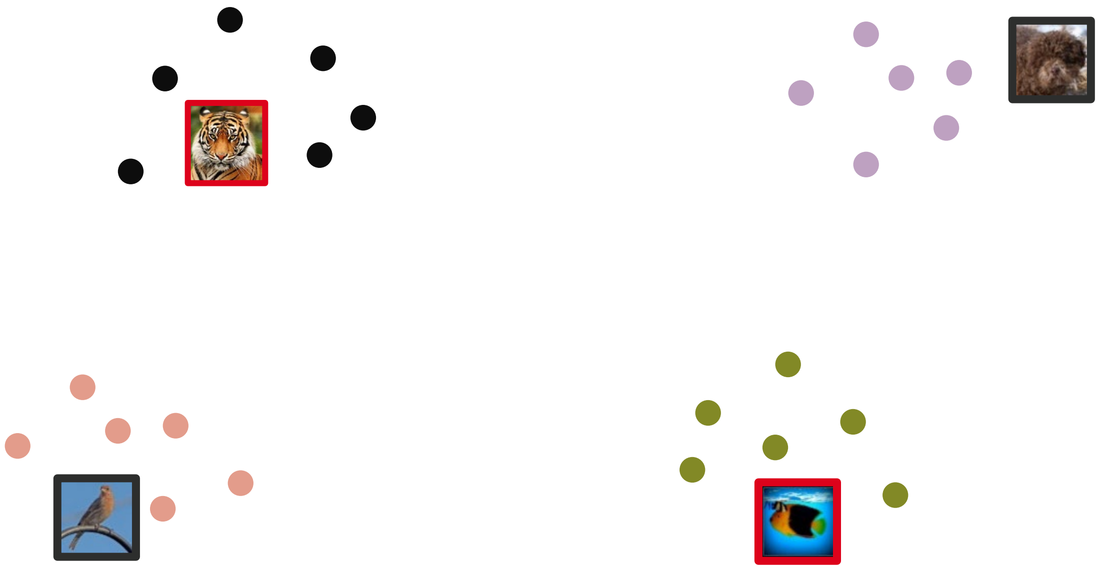
模型遗忘与特征表示



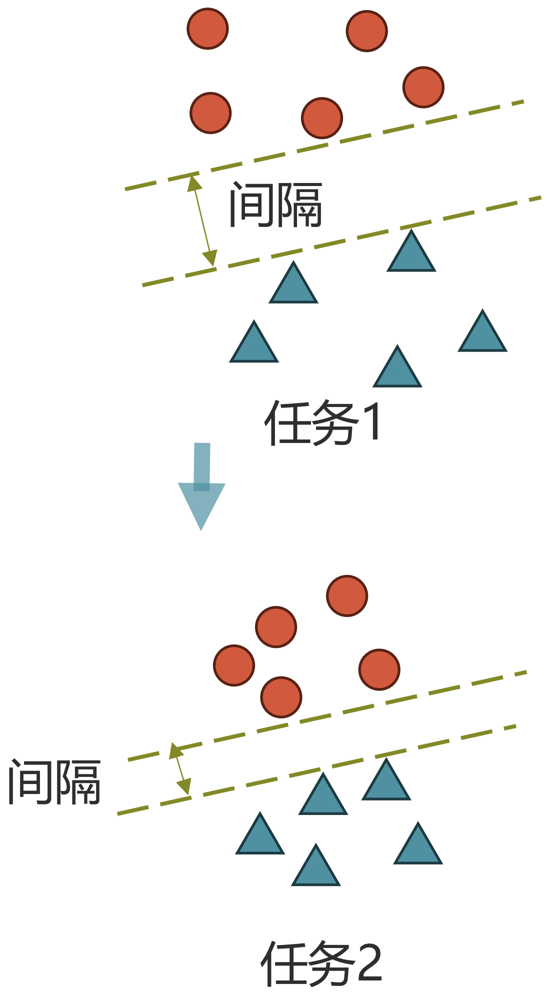




类别增加时，旧类特征表示发生**语义偏移** [Yu et al. CVPR'20]



模型遗忘与特征表示的深入分析



- 特征表示**间隔（margin）**的变化对模型能力的影响

单层间隔 → 全层间隔：全层间隔是在所有层能改变网络分类结果的扰动的最小平方和

$$m_F(\boldsymbol{x}_i, y_i) := \min_{\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_L} \sqrt{\sum_{l=1}^{L} \|\boldsymbol{\delta}_l\|_2^2}$$

$$\text{s.t. } \arg\max F(\boldsymbol{x}_i, \boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_L) \neq y_i$$

泛化界：所有在训练集达到零损失的模型能以$1-\delta$概率满足下式：（即期望损失有关于间隔的上界）

$$\mathbb{E}_P[\ell_{0-1}(F(\boldsymbol{x}), y)] \lesssim \frac{\sum_i \mathcal{C}_i}{\sqrt{n}} \sqrt{\mathbb{E}_{(\boldsymbol{x},y)\sim P_n}\left[\frac{1}{\boxed{m_F(\boldsymbol{x}, y)^2}}\right] \log^2 n + \zeta}$$

提升模型的抗扰动能力

模型更新中间隔与表示空间的变化

全间隔 → 每一层的间隔:

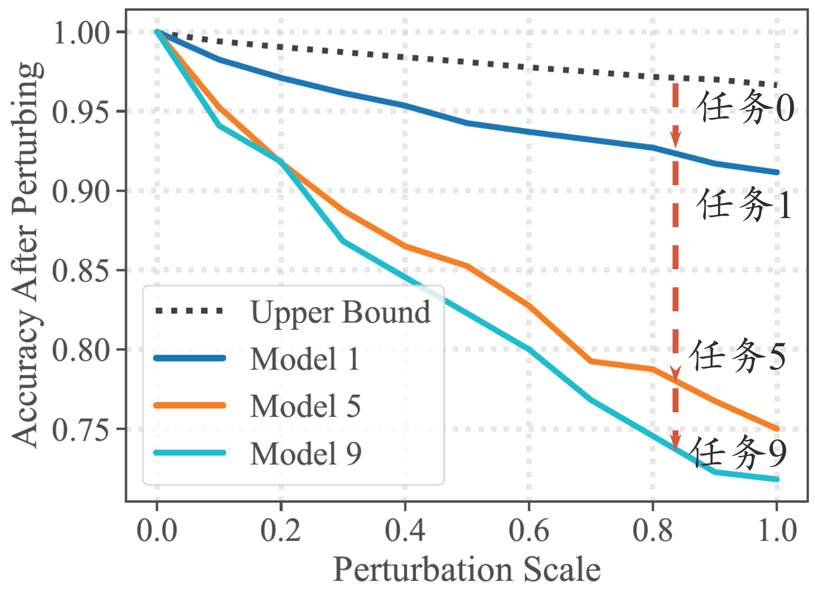
$$m_F(\boldsymbol{x}_i, y_i) \le \tilde{m}_{F,l}(\boldsymbol{x}_i, y_i) := \min_{\boldsymbol{\delta}_l} \|\delta_l\|_2,$$
$$\text{s.t. } \operatorname{argmax} F(\boldsymbol{x}_i, \mathbf{0}, \dots, \delta_l, \dots, \mathbf{0}) \ne y_i.$$

最小扰动的方向 → 模型关于特征的梯度:

$$\tilde{m}_{F,l}(\boldsymbol{x}_i, y_i) = \min_{\delta_l} \|\delta_l\|_2$$
$$\approx \min_{\alpha_{s,l}} \left\|\alpha_{i,l} \nabla_{\boldsymbol{z}} \ell\left(F_l(\boldsymbol{z}_{i,l})\right)\right\|_2$$
$$\text{s.t. } \arg\max F(x_i, 0, \dots, \delta_l, \dots, 0) \ne y_i,$$

$$\tilde{m}_{F,l}(\boldsymbol{x}_i, y_i) \approx \boxed{\alpha_{i,l}^{*} \left\|\nabla_{\boldsymbol{z}} \ell\left(F_l(\boldsymbol{z}_{i,l})\right)\right\|_2}$$

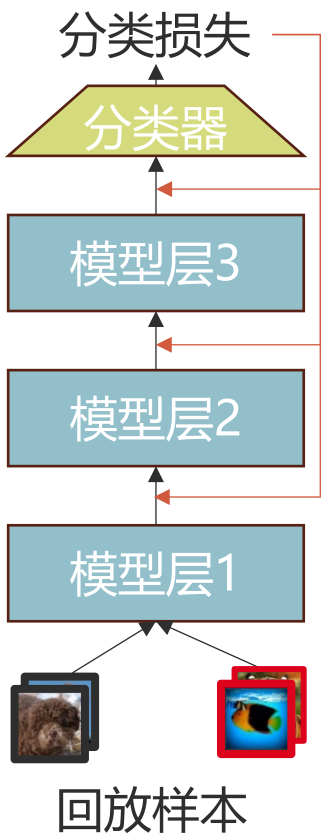
在回放样本上考察扰动强度与预测改变的样本数的关系指示间隔的大小



- 随着模型持续学习，回放样本的抗扰动能力越来越差
- 回放样本的平均间隔在不同层都变得越来越小

Zheng, Zhou, Ye, Zhan. *Multi-layer Rehearsal Feature Augmentation for Class-Incremental Learning*. ICML 2024.

多层回放特征增强



$$\hat{z}_{i,l} = z_{i,l} + \lambda \nabla_{\boldsymbol{z}} \ell(F_l(z_{i,l})),$$

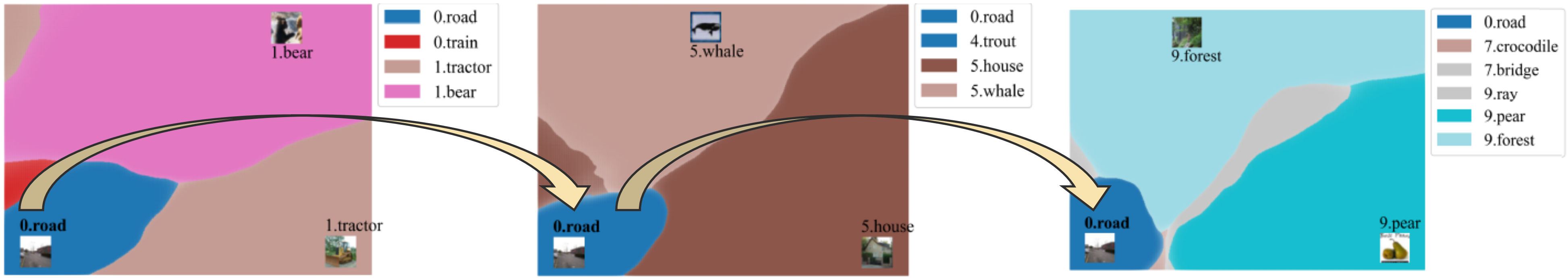
Algorithm 1 Multi-layer Rehearsal Feature Augmentation

Input: batch $\mathcal{B}$, current model $F_t(\boldsymbol{x})$
for $\boldsymbol{x}_i, y_i \in \mathcal{B}$ **do**
  if $\boldsymbol{x}_i, y_i \in \mathcal{M}$ **then** 仅对回放样本进行特征增强
    Sample $l \sim \mathcal{U}\{1, L\}$, $\hat{\beta} \sim \mathcal{U}(0, \beta)$
    $\mathcal{L}_{\text{cls},i} = \text{AugmentedForward}(F_t(\boldsymbol{x}), \boldsymbol{x}_i, y_i, l, \hat{\beta})$
  else
    $\mathcal{L}_{\text{cls},i} = \ell(F_t(\boldsymbol{x}_i), y_i)$
  end if
end for
$\mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{B}|} \sum_i \mathcal{L}_{\text{cls},i}$
Output: $\mathcal{L}_{\text{cls}}$
function AugmentedForward$(F(\boldsymbol{x}), \boldsymbol{x}, y, l, \hat{\beta})$
  $\boldsymbol{z}_l = f_{l-1} \circ \cdots \circ f_1(\boldsymbol{x})$ 用梯度决定增强方向
  $\hat{\boldsymbol{z}}_l = \boldsymbol{z}_l + \hat{\beta} \|\boldsymbol{z}_l\|_2 \nabla_{\boldsymbol{z}} \ell(F_l(\boldsymbol{z}_l), y)$ {Eq. 6}
  return $\ell(F_l(\hat{\boldsymbol{z}}_l), y)$ {Eq. 7} 返回增强后的分类损失
end function

- 在每一层对回放样本进行特征增强来增大全层间隔
- 特征增强的方向由模型关于每层的输入特征的梯度决定

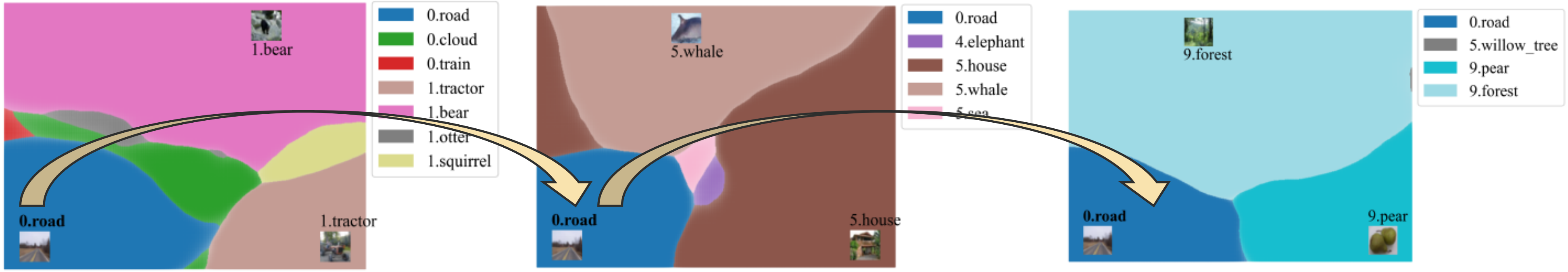
实验结果

使用前



(a) Input Space of Model 1 (b) Input Space of Model 5 (c) Input Space of Model 9

使用后



(a) Input Space of Model 1 (b) Input Space of Model 5 (c) Input Space of Model 9

使用所提出的方法，回放样本的判别边界得到了显著的扩张，表明模型间隔增大

实验结果

Table 1. Performance Experiment Results on CIFAR100. Bold font represents our method improves the baseline in this scenario.

Memory Size	500				1000				2000			
Scenarios	10-10		50-10		10-10		50-10		10-10		50-10	
	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
Replay	30.50	50.83	30.29	41.66	38.55	56.65	38.33	47.72	45.57	61.95	45.80	54.63
w/ MRFA	**31.69**	**51.61**	**31.98**	**42.85**	**39.42**	**57.12**	**39.78**	**48.54**	**46.85**	**62.59**	**47.24**	**55.51**
iCaRL	32.11	53.24	36.16	50.59	41.50	59.98	44.79	56.23	48.65	64.52	50.56	60.08
w/ MRFA	**33.51**	**54.84**	**37.89**	**51.48**	**42.84**	**60.82**	**46.02**	**57.96**	**49.73**	**65.17**	**52.49**	**61.50**
FOSTER	41.54	63.15	48.98	60.32	56.06	71.55	51.40	61.91	62.20	74.49	59.80	67.54
w/ MRFA	**42.12**	**63.90**	**49.51**	**60.83**	**56.76**	**71.94**	**52.06**	**62.34**	**63.41**	**75.23**	**60.74**	**68.02**
DyTox+	52.61	69.29	53.16	65.97	58.47	73.48	56.29	66.71	62.06	75.54	66.75	73.36
w/ MRFA	**54.31**	**70.56**	**54.03**	**66.82**	**59.38**	**74.17**	**57.96**	**67.56**	**63.80**	**76.23**	**68.21**	**74.73**

使用所提出的方法，能够与多种基于样本回放的基线方法进行组合，稳定地提升模型性能

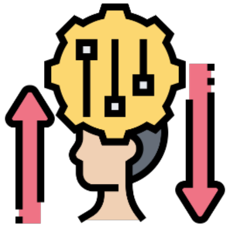
初步结论

在持续学习过程中，模型的表示空间结构难以维持，造成模型能力受损

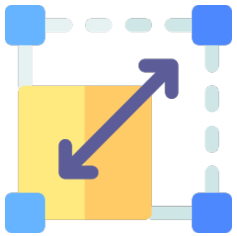
特征表示兼容

特征表示预留

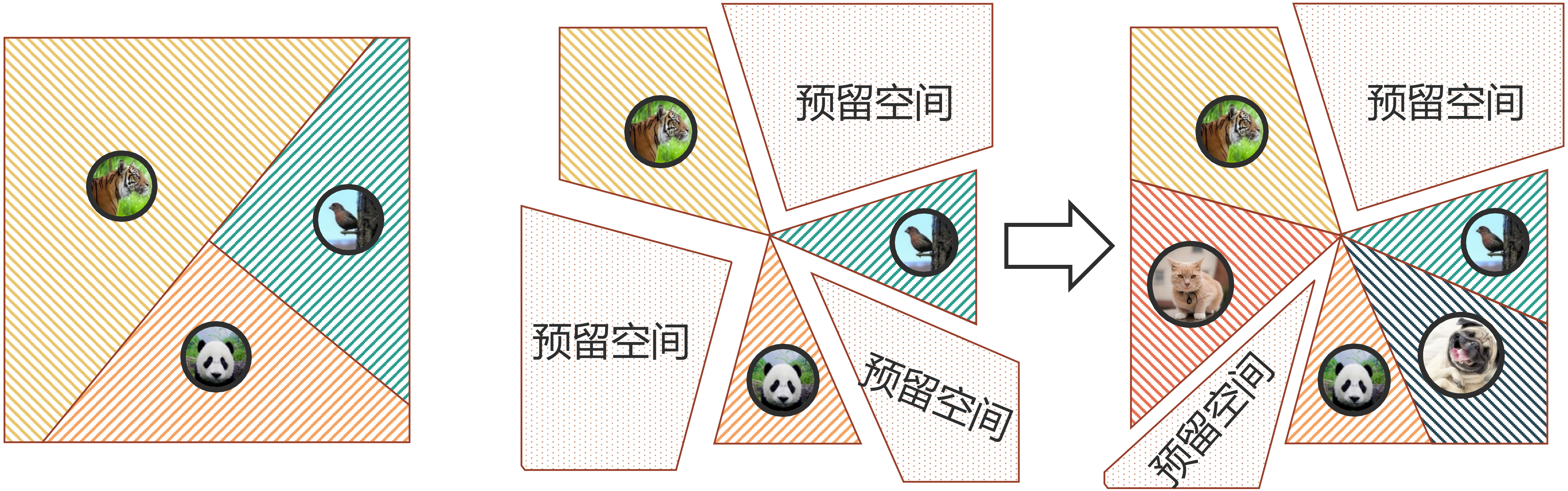
特征表示矫正



特征表示扩张



为新数据预留特征空间



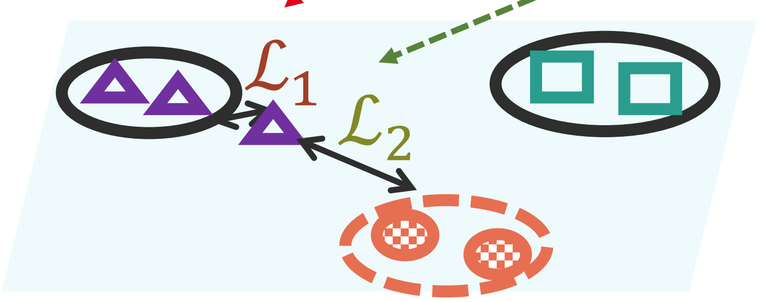
传统训练模式　　　　向前兼容训练模式

Zhou et al., *Forward compatible few-shot class-incremental learning*. CVPR 2022

向前兼容用于小样本增量学习

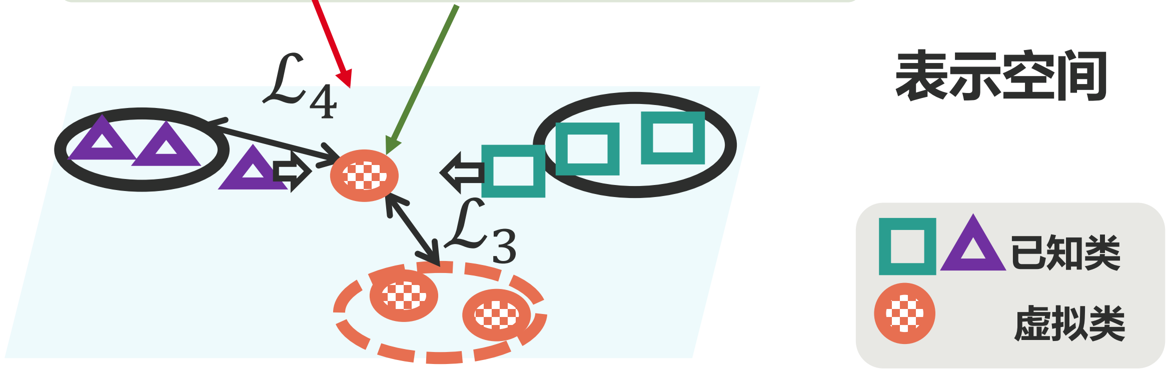
- 核心思想: 为新类预留表示空间

$$\mathcal{L}_v(\mathbf{x}, y) = \underbrace{\ell(f_v(\mathbf{x}), y)}_{\mathcal{L}_1} + \gamma \underbrace{\ell\left(\mathrm{Mask}(f_v(\mathbf{x}), y), \hat{y}\right)}_{\mathcal{L}_2}$$



$$\mathcal{L}_f(\boldsymbol{z}) = \underbrace{\ell(f_v(\boldsymbol{z}), \hat{y})}_{\mathcal{L}_3} + \gamma \underbrace{\ell\left(\mathrm{Mask}(f_v(\boldsymbol{z}), \hat{y}), \hat{\hat{y}}\right)}_{\mathcal{L}_4},$$

$$\boldsymbol{z} = g\left[\lambda h(\mathbf{x}_i) + (1-\lambda) h(\mathbf{x}_j)\right],$$

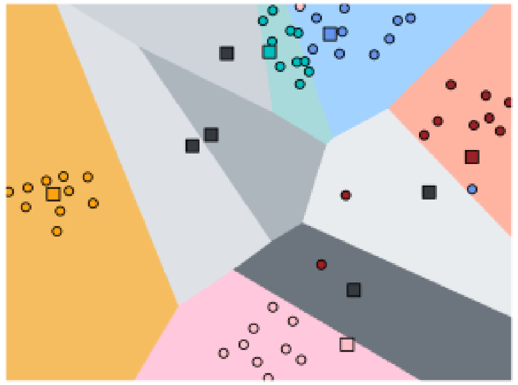


已知类别样本距对应类中心最近、虚拟类中心次近

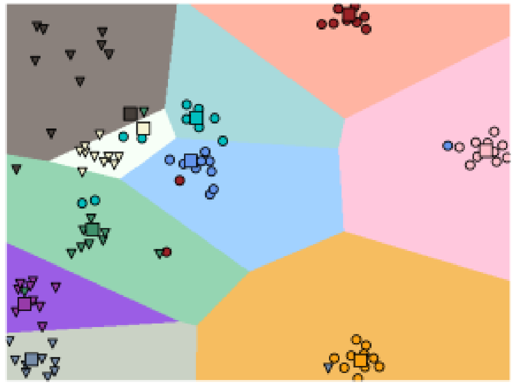
虚拟类样本距对应的虚拟类中心最近、最近的已知类中心次近

特征表示预留的实验验证

Method	Accuracy in each session (%) ↑											PD ↓	Δ PD
	0	1	2	3	4	5	6	7	8	9	10		
Finetune	68.68	43.70	25.05	17.72	18.08	16.95	15.10	10.06	8.93	8.93	8.47	60.21	**+41.25**
Pre-Allocated RPC† [32]	68.47	51.00	45.42	40.76	35.90	33.18	27.23	24.24	21.18	17.34	16.20	52.27	**+33.31**
iCaRL [33]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	47.52	**+28.56**
EEIL [8]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11	46.57	**+27.61**
Rebalancing [21]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87	48.81	**+29.85**
TOPIC [41]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.26	42.40	**+23.44**
SPPR [67]	68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33	31.35	**+12.39**
Decoupled-NegCosine† [26]	74.96	70.57	66.62	61.32	60.09	56.06	55.03	52.78	51.50	50.08	48.47	26.49	**+7.53**
Decoupled-Cosine [45]	75.52	70.95	66.46	61.20	60.86	56.88	55.40	53.49	51.94	50.93	49.31	26.21	**+7.25**
Decoupled-DeepEMD [57]	75.35	70.69	66.68	62.34	59.76	56.54	54.61	52.52	50.73	49.20	47.60	27.75	**+8.79**
CEC [58]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	23.57	**+4.61**
FACT	**75.90**	**73.23**	**70.84**	**66.13**	**65.56**	**62.15**	**61.74**	**59.83**	**58.41**	**57.89**	**56.94**	**18.96**	



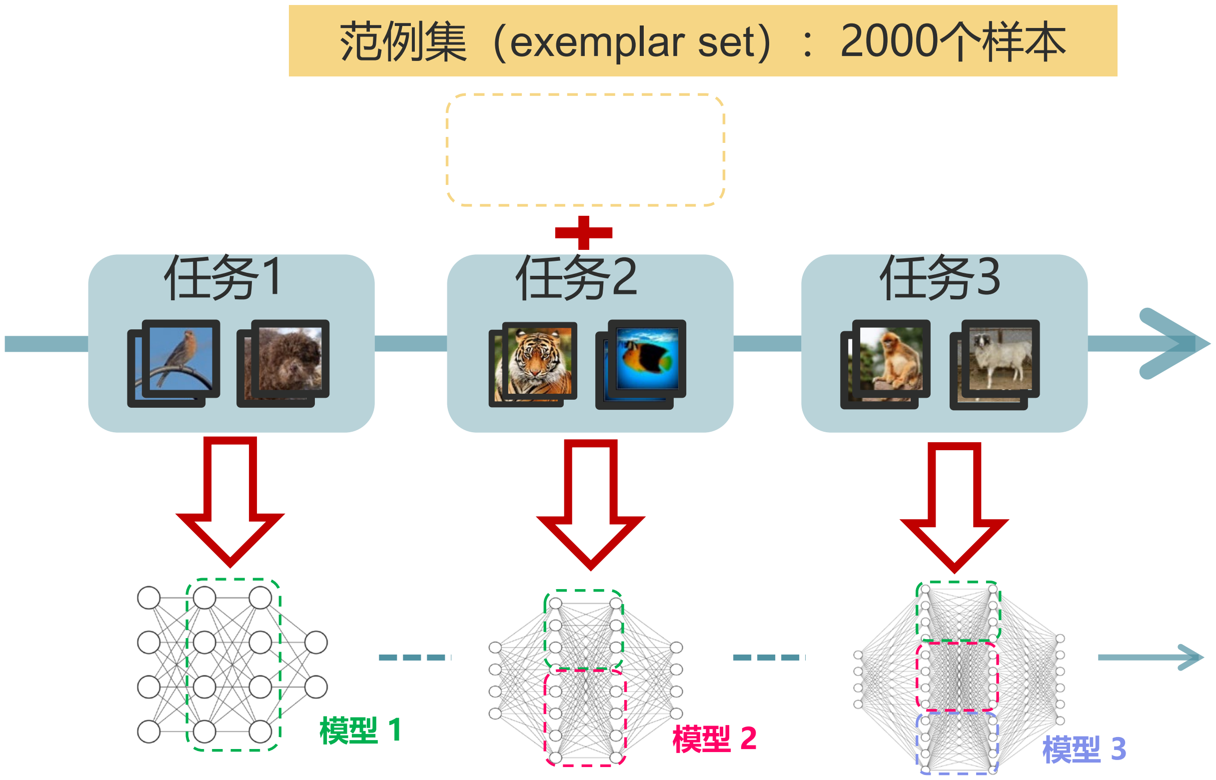
(a) Base session, 5 old classes & 5 virtual prototypes.



(b) Incremental session, 5 old classes & 5 new classes.

- CUB200数据集以100个类别作为base task, 其余类别分10阶段到来的10-way-5-shot设定中, 性能超越SOTA算法约4.5%
- **预留**的新类空间（深色）较好地为模型后续更新提供帮助.

表示空间扩张实现模型兼容



➢ 模型保留、特征拼接 [Yan et al. CVPR’21] [Wang et al. ECCV’22] [Wang et al. NeurIPS’22]

$$f(x) = W^{\top}\mathrm{Concat}[\phi_1(x), \phi_2(x), \ldots \phi_b(x)]$$

保存、冻结旧模型，仅训练新模型

基于范例集进行多个模型之间的预测矫正

资源受限：表示扩张的难题

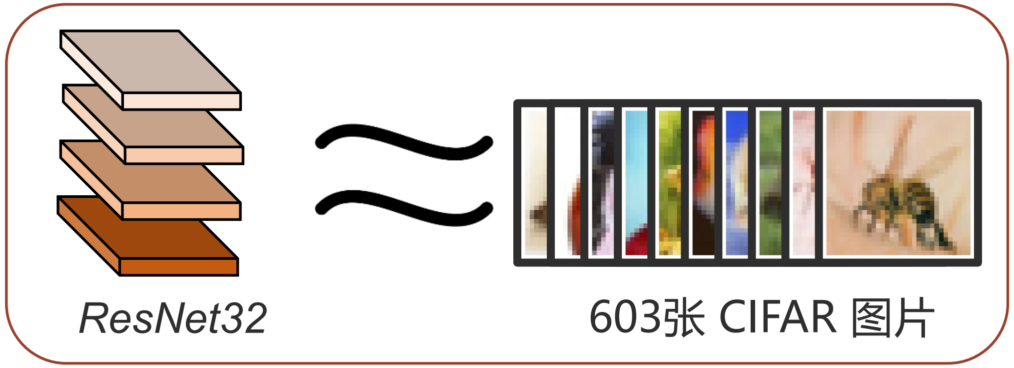
不公平对比

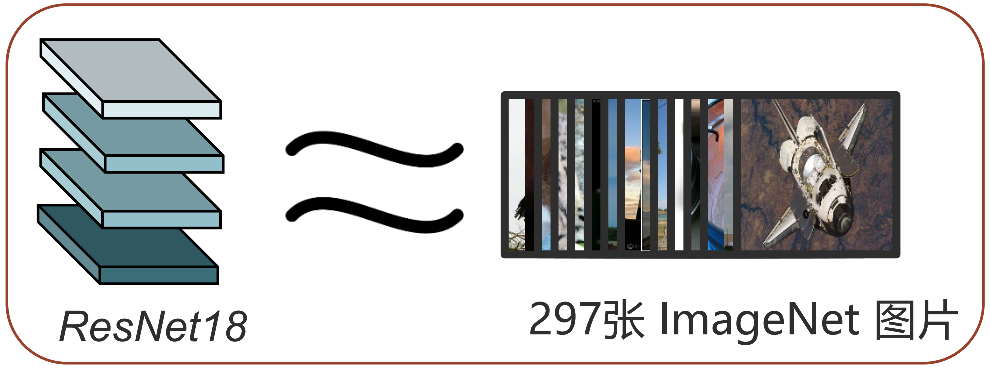
基于回放的方法	2000 范例	模型	
基于模型的方法	2000 范例	模型 * N	

当前增量学习对比协议

基于回放的方法	2000 范例	*Δ* exemplars	模型
基于模型的方法	2000 范例	模型 * N	

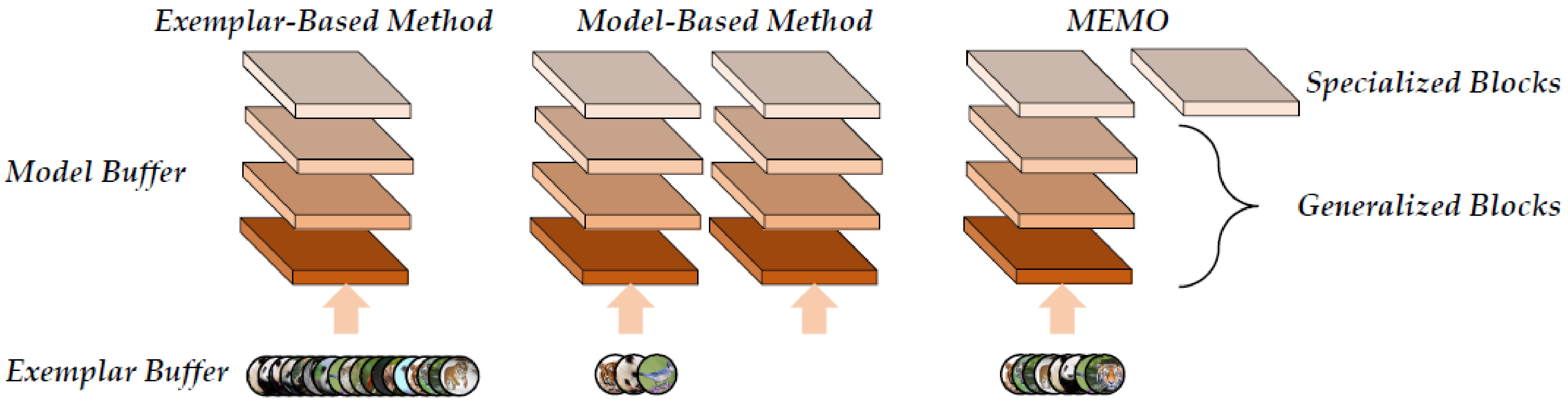
公平对比协议



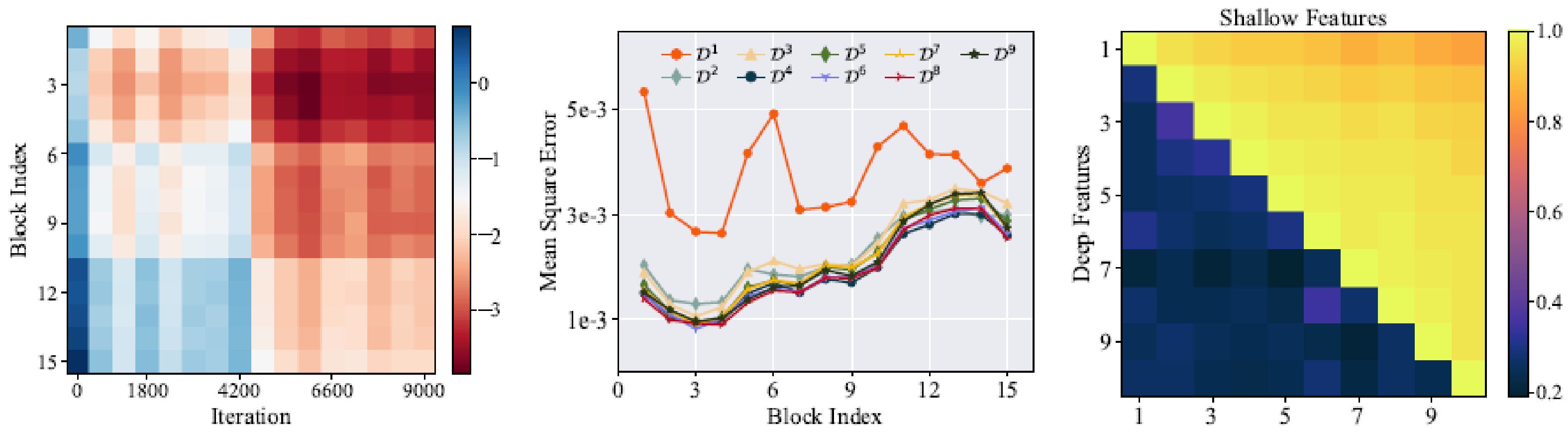


给定**相同的总存储空间**，如何合理分配用于存储数据与模型的空间使其更好复用表示

局部模型的兼容



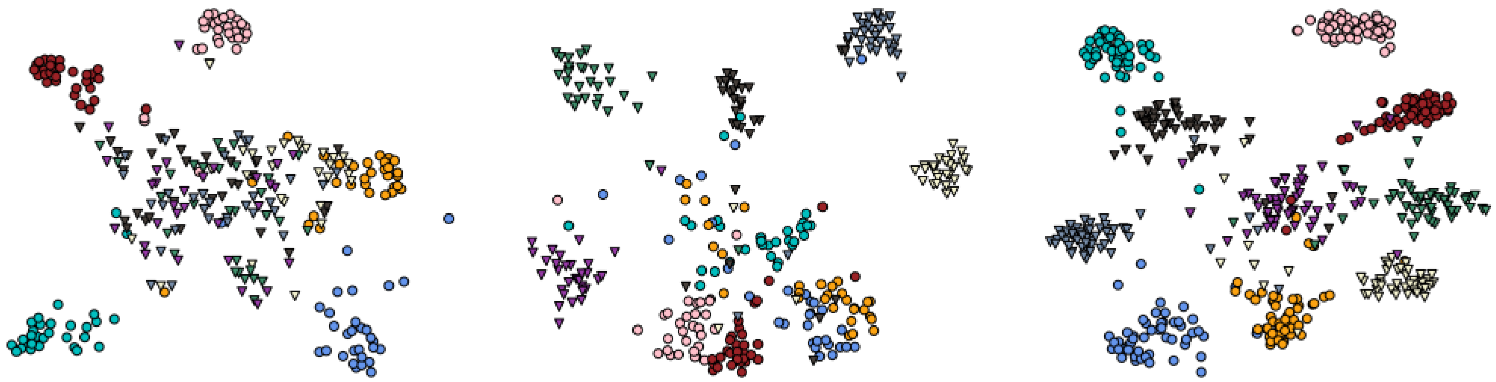
- 给定模型存储空间,如何合理分配用于存储数据与模型的空间使其更好复用表示



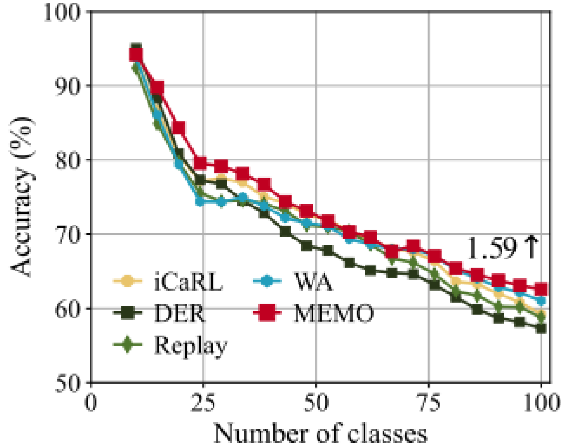
(a) Gradient norm (log scale) (b) MSE of different blocks (c) CKA between backbones

Zhou, et al., *A Model or 603 Exemplars: Towards Memory-Efficient Class-Incremental Learning*. ICLR 2023.

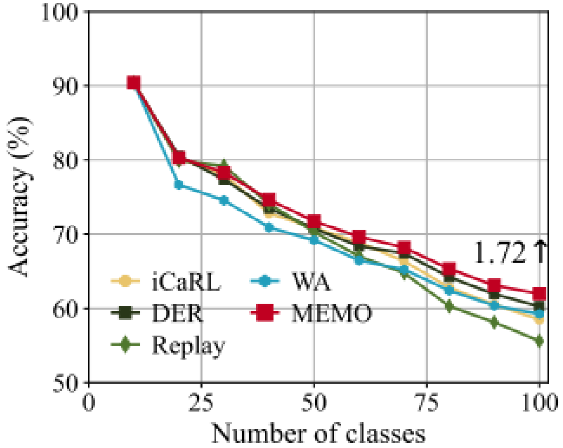
实验验证



(a) $\phi_{s1}(\phi_g(\mathbf{x}))$ (b) $\phi_{s2}(\phi_g(\mathbf{x}))$ (c) $[\phi_{s1}(\phi_g(\mathbf{x})), \phi_{s2}(\phi_g(\mathbf{x}))]$



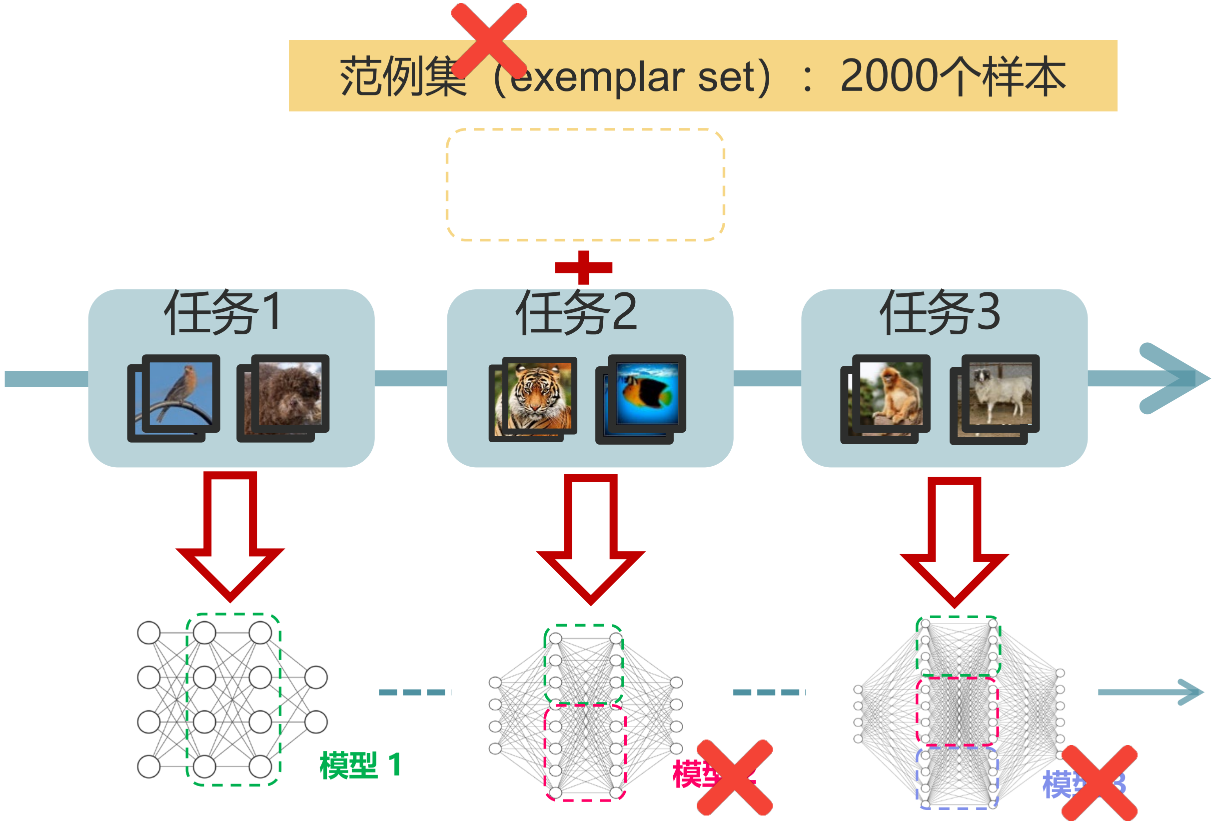
(a) CIFAR100 Base0 Inc5



(b) CIFAR100 Base0 Inc10

- 在共用浅层特征的同时, 深层特征学得了**任务相关**的特征表示;
- 将不同任务的深层特征合并时, 能够获得**适应所有类别**的特征表示.
- 当所有算法存储开销相同时, 所提出算法实现了免费的性能提升.

不基于回放样本的表示空间扩张



范例集的使用也增加了算法对资源的消耗，若不使用范例集，则算法无法进行预测校准

若以预训练模型作为初始化，由于预训练模型使用复杂的网络结构，保存**多个主干网络**将显著消耗大量存储空间

怎样在仅使用**固定数目主干网络**的同时，**不使用回放样本**进行表示空间扩张?

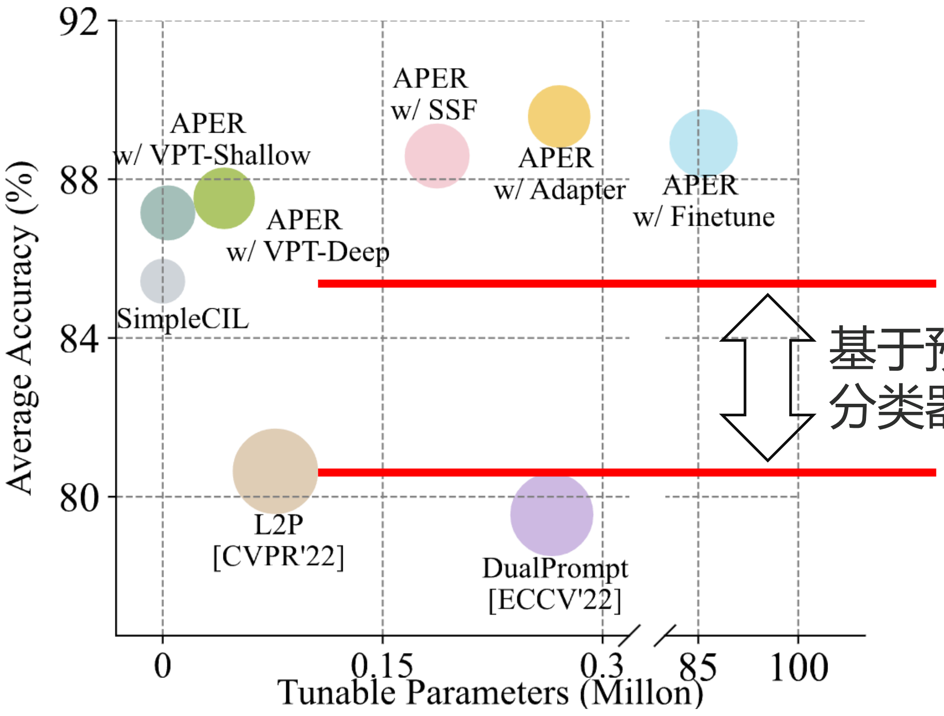
基于视觉预训练模型的解决思路

- 增量学习的目的是获得适配所有任务的特征表示，并抵抗学习过程中的特征遗忘
- 相比于传统的从零开始训练设定，预训练模型天然具有**可泛化**的特征表示

预训练模型能为特征表示复用带来何种便捷?

“从零训练”　　“基于预训练模型”

$$\boldsymbol{p}_i=\frac{1}{K}\sum_{j=1}^{|\mathcal{D}^b|}\mathbb{I}(y_j=i)\phi(\mathbf{x}_j)$$



基于预训练模型的原型分类器轻松超过SOTA

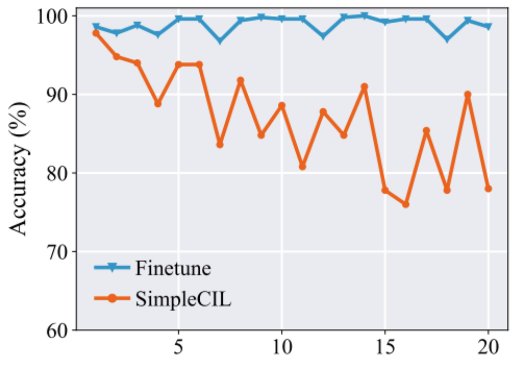
基于预训练模型是否需要增量学习?

Zhou, et al., *Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need*. IJCV 2025.

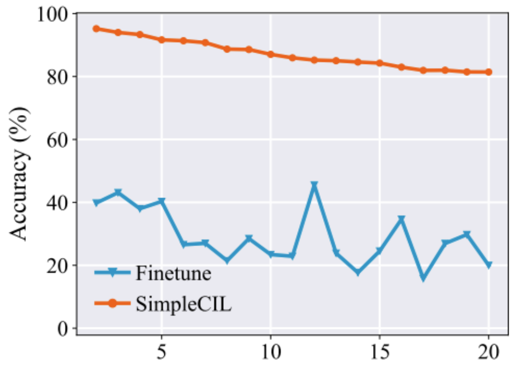
基于视觉预训练模型的解决思路

（预训练模型+原型分类器）是否足以处理
任何下游任务的增量学习?

不能！利用下游任务调整可进一步提升模型的能力

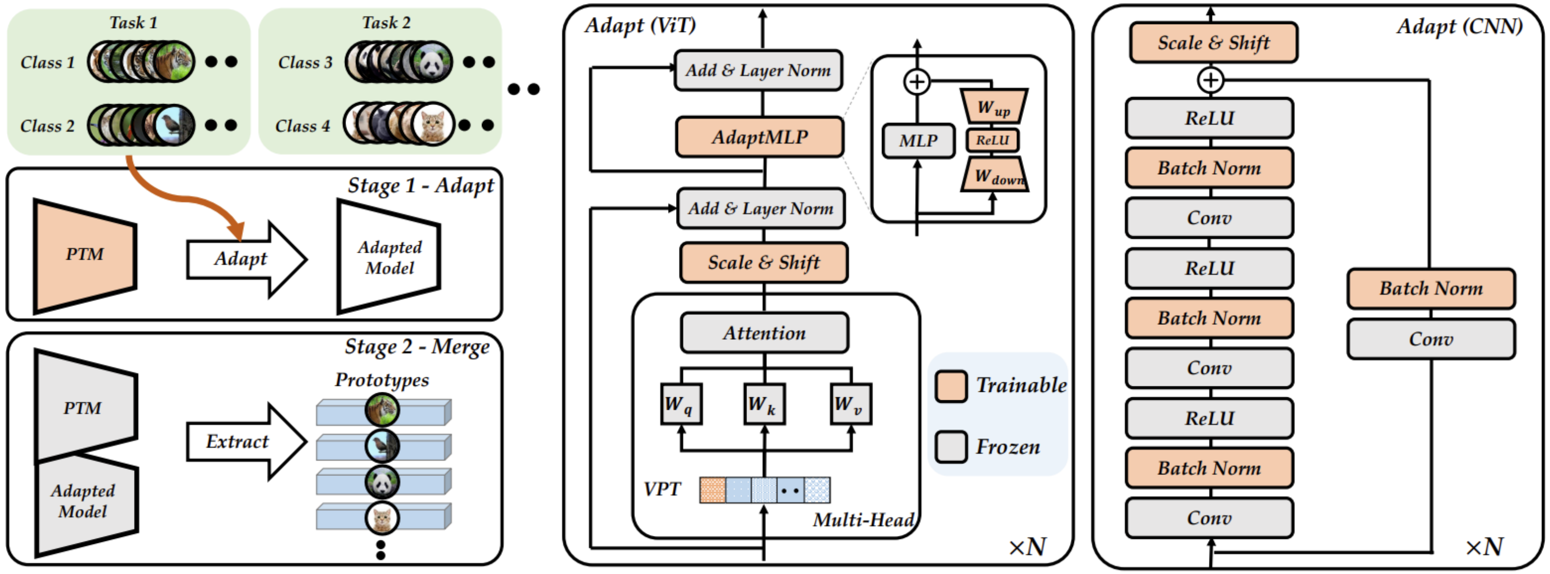


新类性能



旧类性能

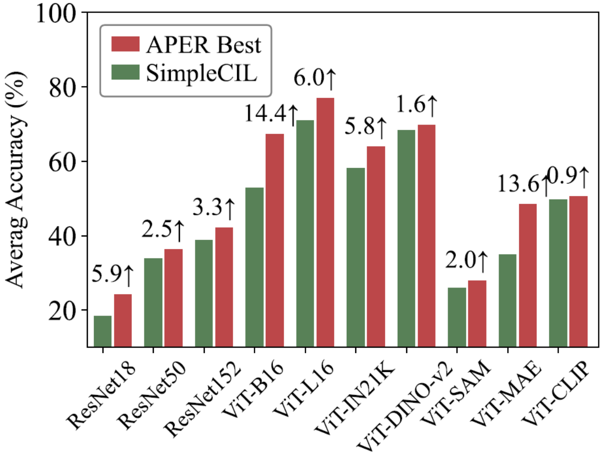
如何结合预训练模型与适配后模型的优势?

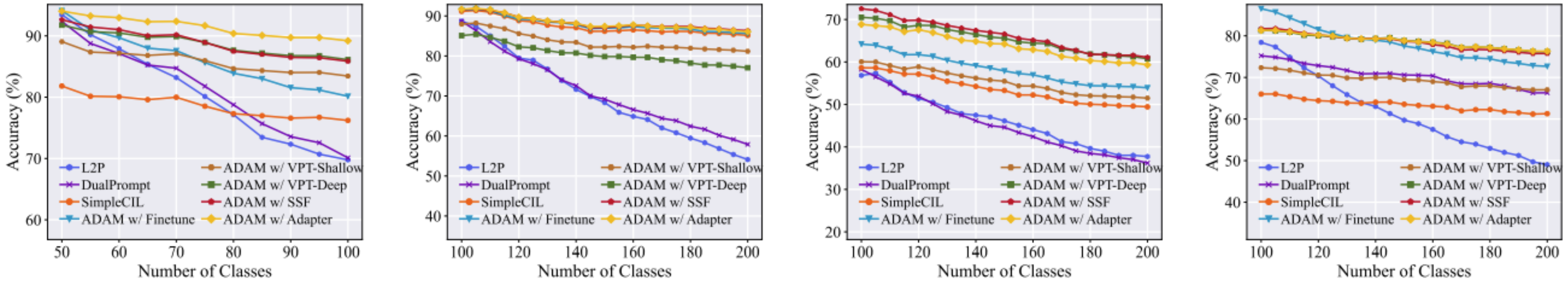


第一阶段：模型适配与拼接
后续阶段：原型分类器

实验验证

Method	CIFAR B0 Inc5 $\bar{\mathcal{A}}$	CIFAR B0 Inc5 $\mathcal{A}_B$	CUB B0 Inc10 $\bar{\mathcal{A}}$	CUB B0 Inc10 $\mathcal{A}_B$	IN-R B0 Inc5 $\bar{\mathcal{A}}$	IN-R B0 Inc5 $\mathcal{A}_B$	IN-A B0 Inc10 $\bar{\mathcal{A}}$	IN-A B0 Inc10 $\mathcal{A}_B$	ObjNet B0 Inc10 $\bar{\mathcal{A}}$	ObjNet B0 Inc10 $\mathcal{A}_B$	OmniBench B0 Inc30 $\bar{\mathcal{A}}$	OmniBench B0 Inc30 $\mathcal{A}_B$	VTAB B0 Inc10 $\bar{\mathcal{A}}$	VTAB B0 Inc10 $\mathcal{A}_B$
Finetune	38.90	20.17	26.08	13.96	21.61	10.79	21.60	10.96	19.14	8.73	23.61	10.57	34.95	21.25
Finetune Adapter [10]	60.51	49.32	66.84	52.99	47.59	40.28	43.05	37.66	50.22	35.95	62.32	50.53	48.91	45.12
LwF [54]	46.29	41.07	48.97	32.03	39.93	26.47	35.39	23.83	33.01	20.65	47.14	33.95	40.48	27.54
SDC [111]	68.21	63.05	70.62	66.37	52.17	49.20	26.65	23.57	39.04	29.06	60.94	50.28	45.06	22.50
L2P [101]	85.94	79.93	67.05	56.25	66.53	59.22	47.16	38.48	63.78	52.19	73.36	64.69	77.11	77.10
DualPrompt [100]	87.87	81.15	77.47	66.54	63.31	55.22	52.56	42.68	59.27	49.33	73.92	65.52	83.36	81.23
CODA-Prompt [82]	89.11	81.96	84.00	73.37	64.42	55.08	48.51	36.47	66.07	53.29	77.03	68.09	83.90	83.02
CPP [55]	85.21	78.64	86.60	85.27	64.33	60.74	53.70	40.70	60.44	49.92	71.52	73.26	85.92	84.30
LAE [24]	**92.47**	**87.62**	83.13	77.78	69.05	63.17	57.19	46.41	62.28	50.57	73.80	70.63	86.14	84.39
SimpleCIL	87.57	81.26	92.20	**86.73**	62.58	54.55	60.50	49.44	65.45	53.59	79.34	73.15	85.99	84.38
APER w/ Finetune	87.67	81.27	91.82	86.39	70.51	62.42	61.57	50.76	61.41	48.34	73.02	65.03	**87.47**	80.44
APER w/ VPT-Shallow	90.43	84.57	92.02	86.51	66.63	58.32	57.72	46.15	64.54	52.53	79.63	73.68	87.15	**85.36**
APER w/ VPT-Deep	88.46	82.17	91.02	84.99	68.79	60.48	60.59	48.72	67.83	54.65	**81.05**	**74.47**	86.59	83.06
APER w/ SSF	87.78	81.98	91.72	86.13	68.94	60.60	**62.81**	**51.48**	**69.15**	**56.64**	80.53	74.00	85.66	81.92
APER w/ Adapter	90.65	85.15	**92.21**	**86.73**	**72.35**	**64.33**	60.53	49.57	67.18	55.24	80.75	74.37	85.95	84.35

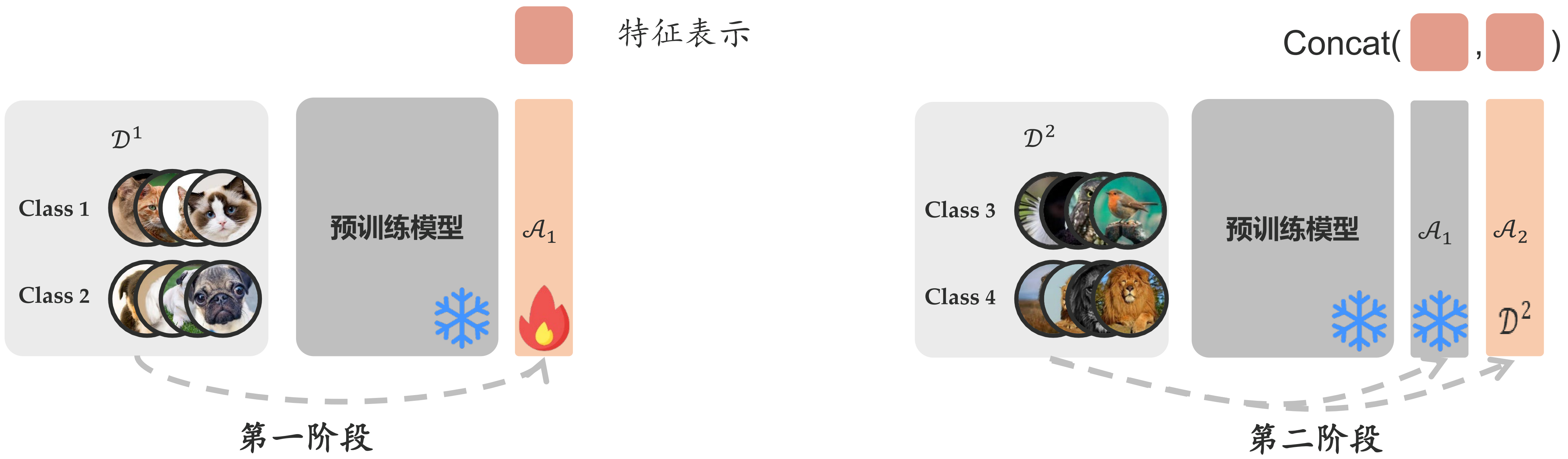




(a) CIFAR B50 Inc5 (b) CUB B100 Inc5 (c) ImageNet-A B100 Inc5 (d) ImageNet-R B100 Inc5

- 在7个数据集不同设定下的增量学习场景中，均对比当前SOTA有稳定性能提升
- 在不同骨干网络上，可以实现稳定性能提升

轻量级特征表示扩张



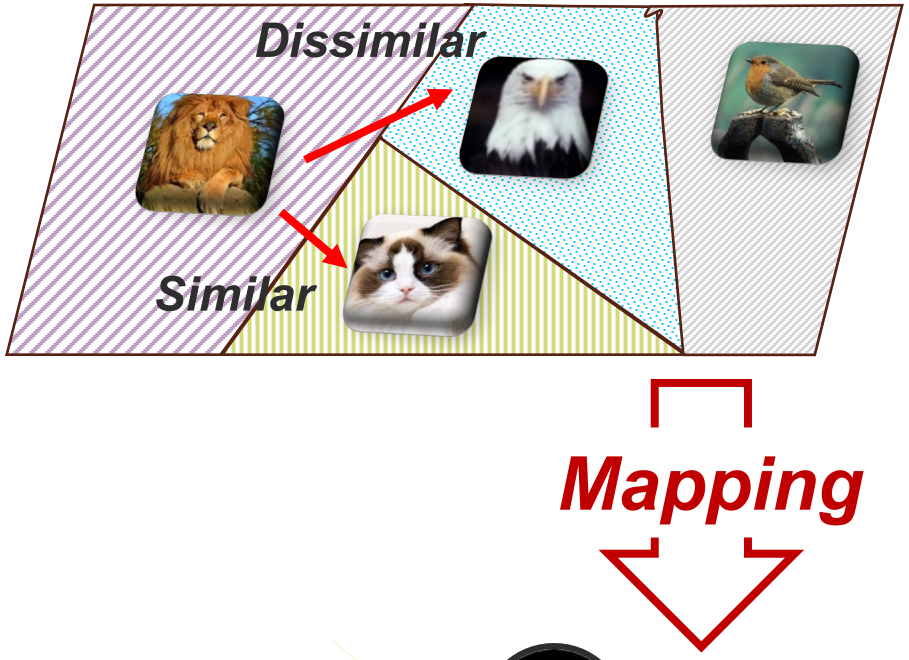
在预训练模型基础上，通过轻量级微调为每个新任务创建**新的特征子空间**，从而可以协同考虑所有子空间进行模型联合决策

由于所有轻量级微调模块共享预训练权重，为模型进行特征表示扩张所需的参数规模**极小**

Zhou et al., *Expandable Subspace Ensemble for Pre-Trained Model-Based Class-Incremental Learning*. CVPR 2024

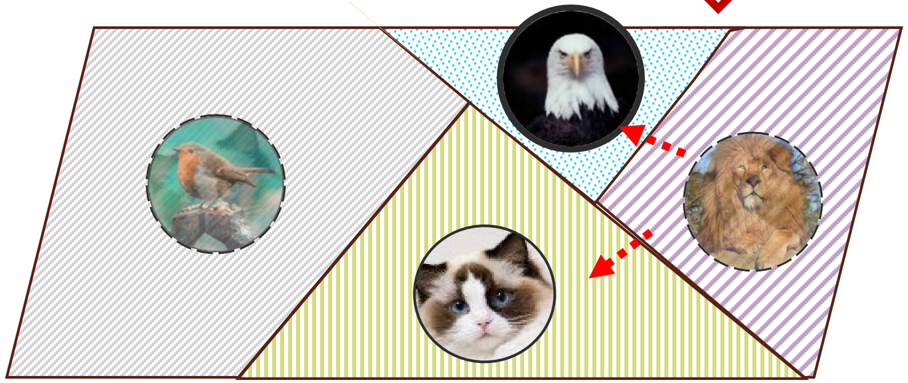
语义引导的原型补全

Subspace of $\mathcal{A}_1$

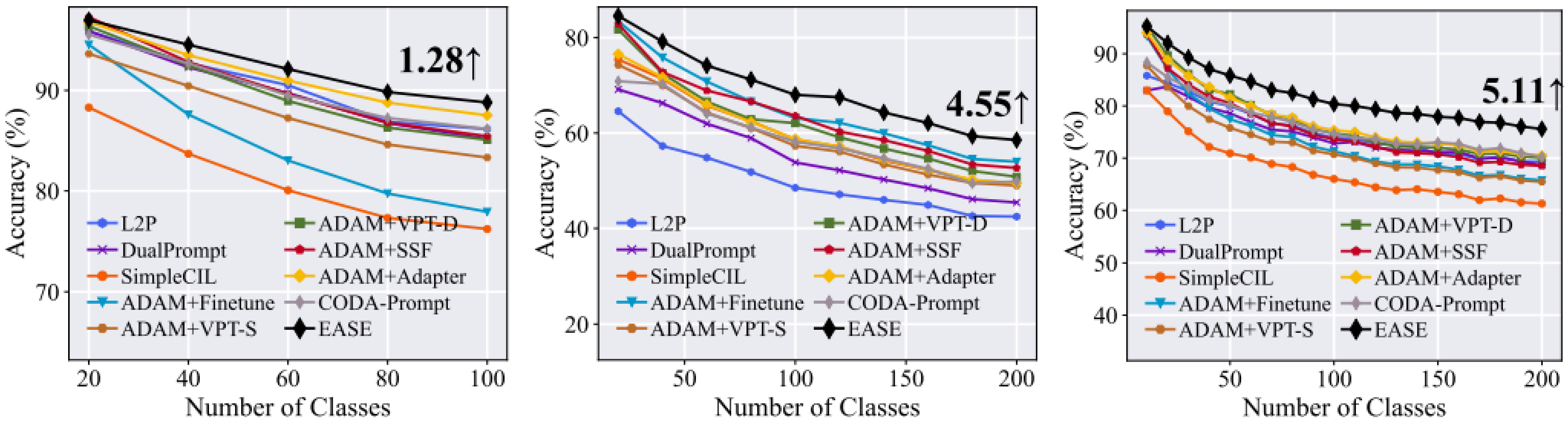


Subspace of $\mathcal{A}_2$

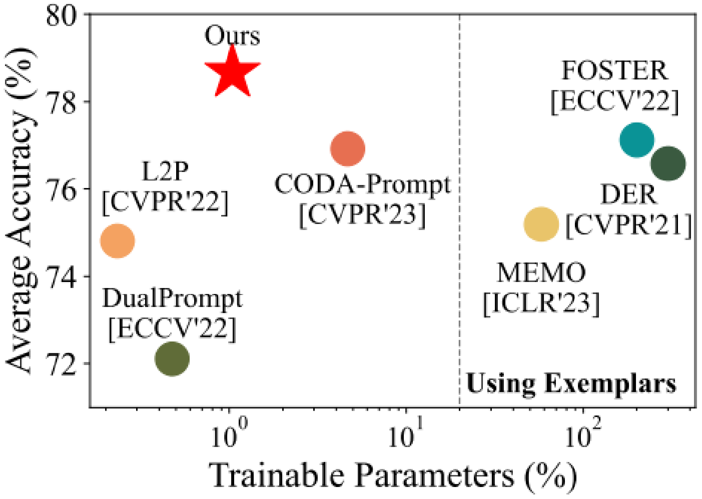
由于不保存旧数据，模型在扩张特征表示时面临（**特征维度-分类器维度）的不匹配**



实验验证



(a) CIFAR B0 Inc20　　(b) ImageNet-A B0 Inc20　　(c) ImageNet-R B0 Inc10



- 在多个基准数据集的实验对比中，相比当前最优方法提升1-5%
- 模型以**极低代价**进行表示空间扩张，能够在不依赖回放样本与大规模参数微调的情况下实现最优性能

多模态大语言模型中的模型兼容

初始化的纯文本LLM　多模态LLM

Model / 纯文本(Text-only)测评数据		Qwen LLM	Qwen + CLIP	LoRAQ + CLIP	Qwen + SigLIP
Exam	MMLU	**60.86**	50.83	59.67	51.16
	CMMLU	<u>69.37</u>	62.58	67.87	60.46
	ARC-E	**59.96**	56.93	<u>59.35</u>	55.87
	ARC-C	38.90	39.14	38.64	<u>39.50</u>
Under-standing	Winogrande	**71.38**	69.82	<u>71.03</u>	69.05
	OpenbookQA	**81.73**	78.31	<u>81.29</u>	77.51
	Race-Middle	**74.82**	68.25	72.06	68.34
	Race-High	**71.05**	59.20	65.67	57.72
	WSC	<u>56.17</u>	54.18	57.30	55.23
	CHID	71.94	71.82	72.92	**74.29**
Reasoning	HellaSwag	**65.70**	61.90	[illegible]	63.24
	SIQA	45.57	<u>50.20</u>	46.83	**51.71**
	PIQA	76.59	74.60	73.77	75.19
Math	GSM8K	56.77	50.10	54.25	<u>51.37</u>
Code	MBPP	<u>37.50</u>	34.82	36.72	33.20

多模态LLM存在纯文本遗忘的问题

Text-only Instruction

User: What are some refreshing fruits to beat the summer heat?

WINGS (Ours): (…) watermelon, strawberries, blueberries, raspberries, peaches, and oranges.

Interleaved Image-Text Context

User: Oh wow! I just found two watermelons in the garden: They're gonna be super refreshing.

WINGS (Ours): The image shows two watermelons, one **labeled** as “1” and the other as “2.” They are placed **next to each other**, and both are ripe. The watermelons are **green** and have **a striped pattern**…

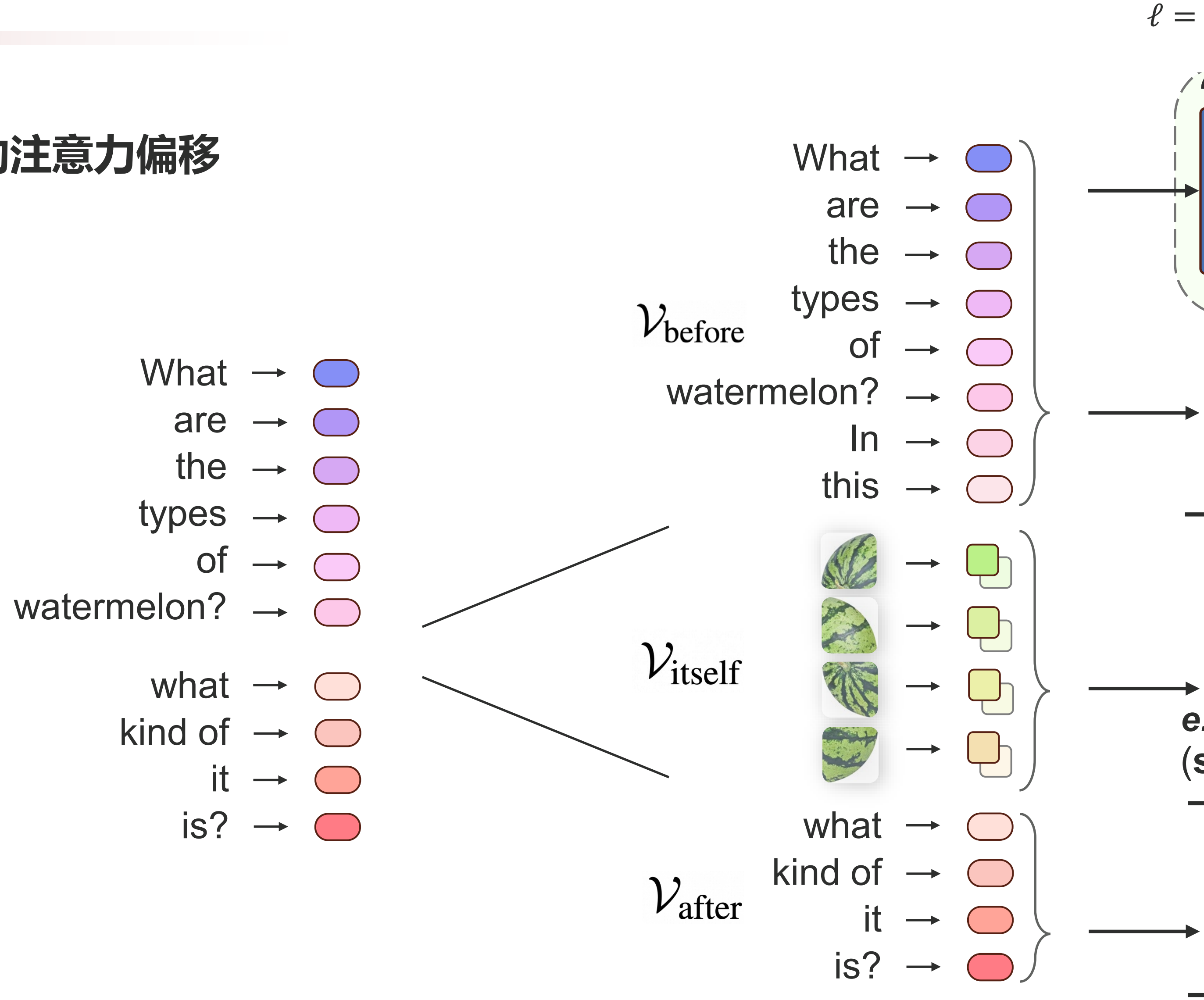
如何在扩展LLM能力的同时，
维持其在文本任务上的能力？

多模态LLM遗忘了已有的纯文本能力

在视觉输入token的增量学习前后发生了较大的注意力偏移

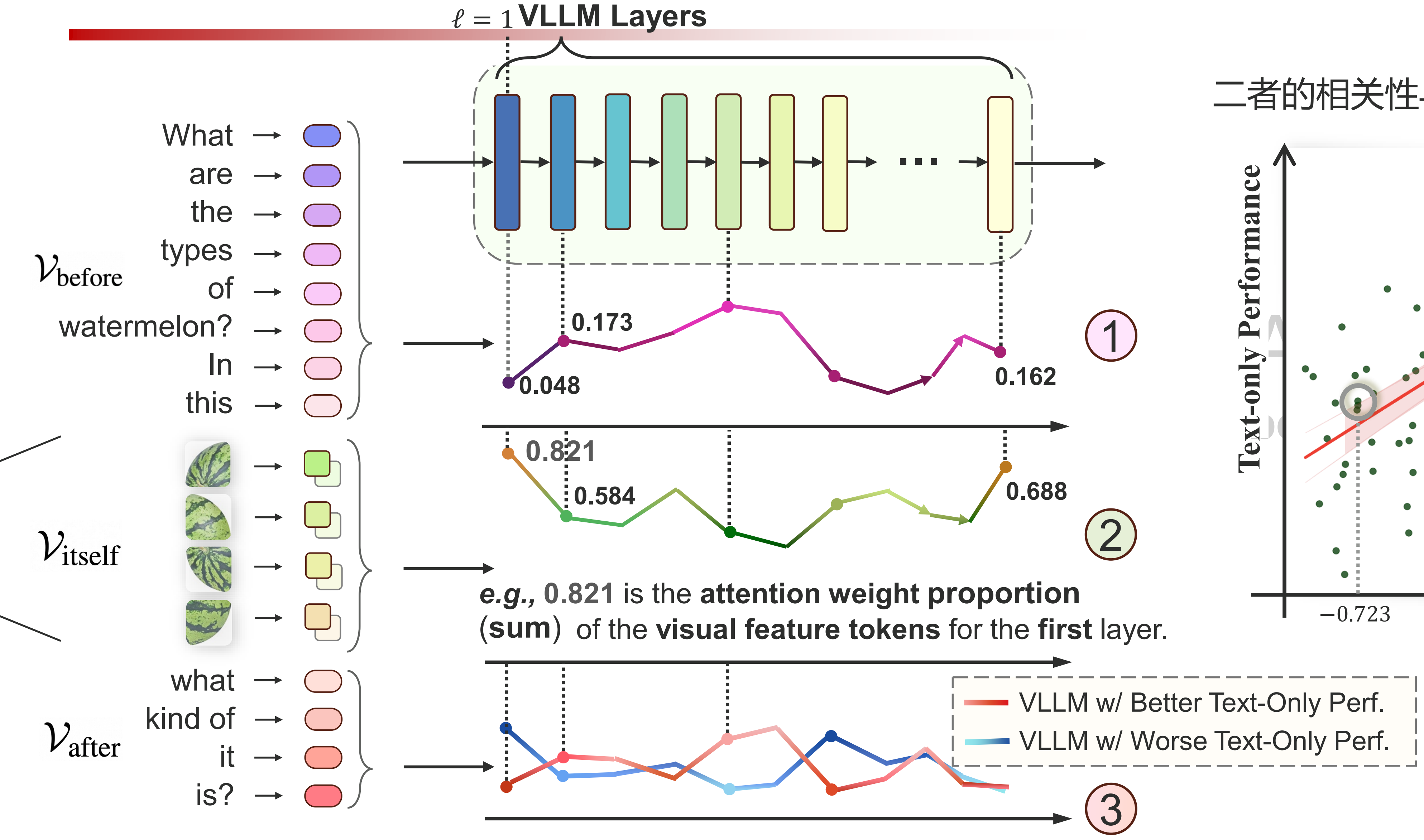
多模态大语言模型

是大语言模型**兼容视觉特征的**
增量（持续）训练

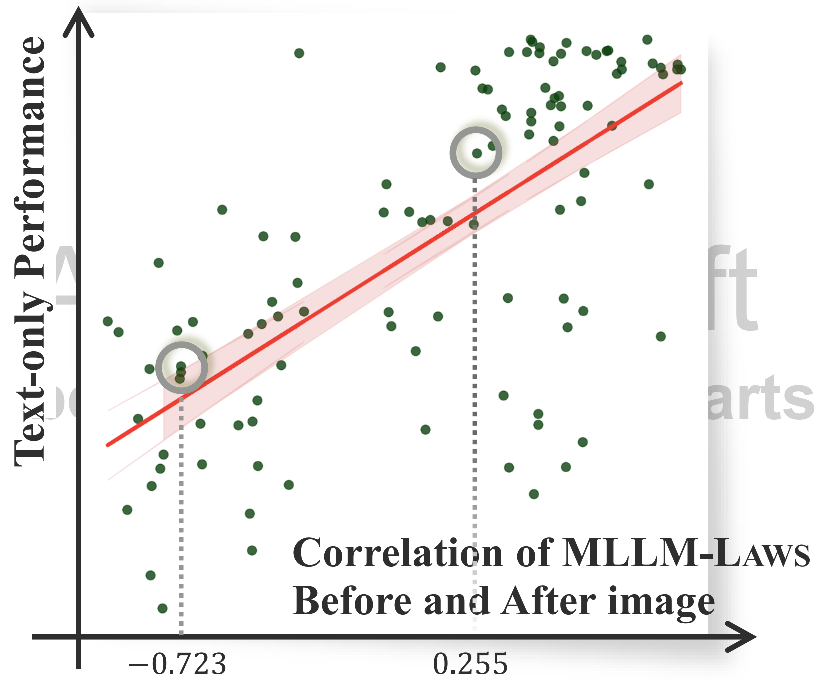


Yi-Kai Zhang, Shiyin Lu, Yang Li, YanQing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, Han-Jia Ye. Wings: Learning Multimodal LLMs without Text-only Forgetting. NeurIPS 2024.

多模态LLM遗忘了已有的纯文本能力



二者的相关性与模型纯文本能力正相关

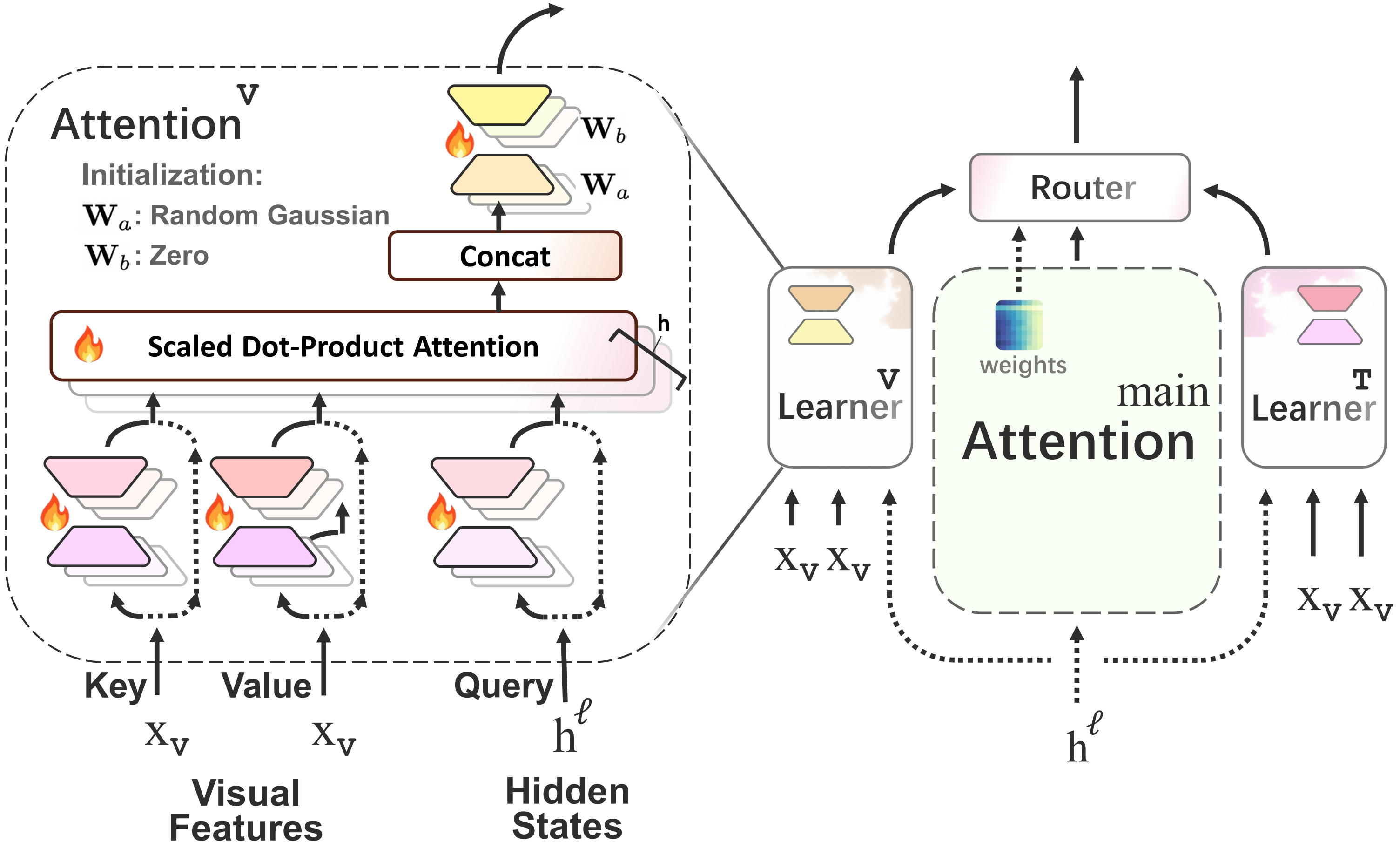


Wings

添加额外的模态专家，矫正结构兼容过程中的注意力偏移

Modality Learners

w/ Low-Rank Residual Attention



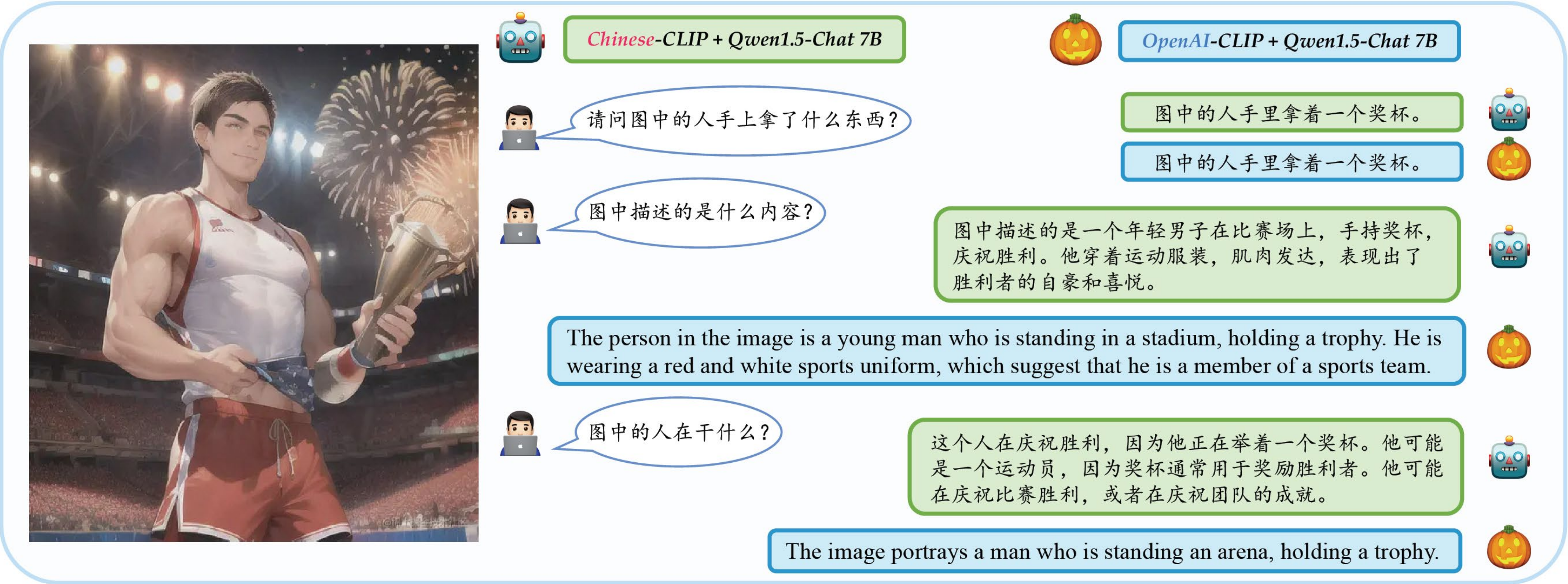
Wings

对比领域领先开源多模态LLM，保持了Text-only和多模态指令上的通用性能

Dataset		Vicuna LLM	Vicuna + CLIP	LoRA$_{Vicu.}$ + CLIP	Vicuna + SigLIP	Qwen LLM	Qwen + CLIP	LoRA$_{Q.}$ + CLIP	Qwen + SigLIP	WINGS (Ours)	Text-only Forgetting (-)	Our Impro. (-)
Exam	MMLU	51.18	51.12	48.89	50.63	**60.86**	50.83	59.67	51.16	60.53	9.70	9.37
	CMMLU	38.60	38.29	37.24	38.73	69.37	62.58	67.87	60.46	**69.82**	8.91	9.36
	ARC-E	57.62	53.63	55.82	53.95	**59.96**	56.93	59.35	55.87	54.29	4.09	-1.58
	ARC-C	33.75	34.60	34.68	35.17	38.90	39.14	38.64	39.50	**43.39**	-0.60	3.89
Under-standing	Winogrande	68.01	64.97	67.83	65.21	**71.38**	69.82	71.03	69.05	69.28	2.33	0.23
	OpenbookQA	77.10	73.28	77.15	72.12	**81.73**	78.31	81.29	77.51	81.05	4.22	3.54
	Race-Middle	63.99	60.10	62.84	59.45	**74.82**	68.25	72.06	68.34	74.24	6.48	5.90
	Race-High	58.74	53.24	54.91	52.69	**71.05**	59.20	65.67	57.72	69.62	13.33	11.90
	WSC	51.30	47.21	51.06	47.72	56.17	54.18	57.30	55.23	**66.35**	0.94	11.12
	CHID	39.05	49.66	45.26	53.49	71.94	71.82	72.92	**74.29**	74.06	-2.35	-0.23
Reasoning	HellaSwag	63.11	63.08	62.58	63.02	**65.70**	61.90	64.32	63.24	65.12	2.46	1.88
	SIQA	42.37	44.06	43.27	44.52	45.57	50.20	46.83	**51.71**	49.64	-6.14	-2.07
	PIQA	71.92	71.95	70.35	71.84	76.59	74.60	73.77	75.19	**78.06**	1.40	2.87
	OCNLI	33.89	37.74	39.41	40.46	49.73	48.31	48.07	50.29	**50.39**	-0.56	0.10
Math	GSM8K	25.19	23.72	22.68	23.05	56.77	50.10	54.25	51.37	**52.08**	5.40	0.71
Code	MBPP	13.80	11.29	13.92	10.80	37.50	34.82	36.72	33.20	**38.92**	4.30	5.72
Multimodal	MMMU-VAL	–	35.67	30.78	35.56	–	34.56	32.33	35.11	**39.89**	–	4.78
	MMMU-TEST	–	34.40	30.90	35.33	–	34.90	31.80	35.10	**37.30**	–	2.20
	MMBench	–	63.18	59.83	65.14	–	66.05	62.84	**70.94**	70.53	–	-0.41
	ScienceQA	–	67.72	64.49	71.50	–	74.26	69.09	74.89	**78.76**	–	3.87

- LLMs: Vicunav1.5 & Qwen1.5
- Visual: CLIP, SigLIP
- Wings: Qwen1.5 + SigLIP

多模态LLM遗忘了已有的多语言能力

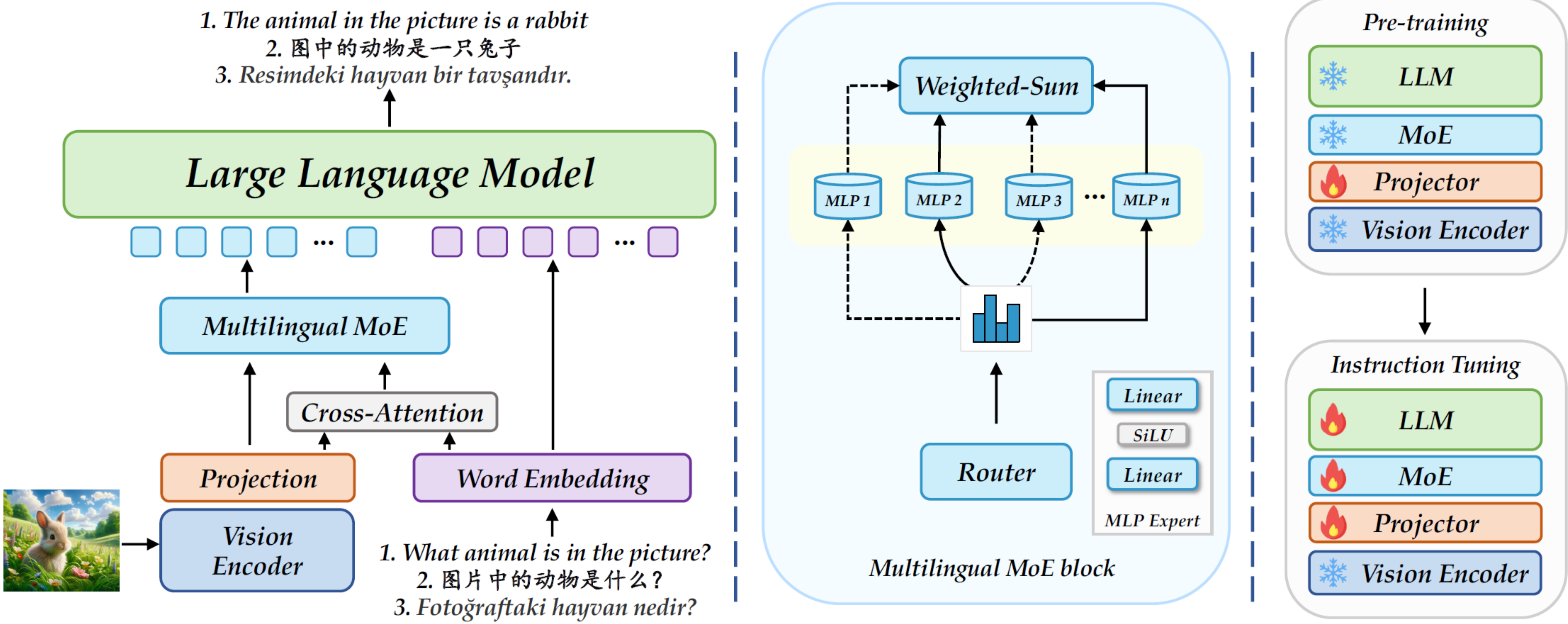


处理多语言任务时，多模态大模型受到Visual Encoder训练时常见语言的影响，丧失LLM中本有的多语言能力

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, Han-Jia Ye. Parrot: Multilingual Visual Instruction Tuning. ICML 2025.

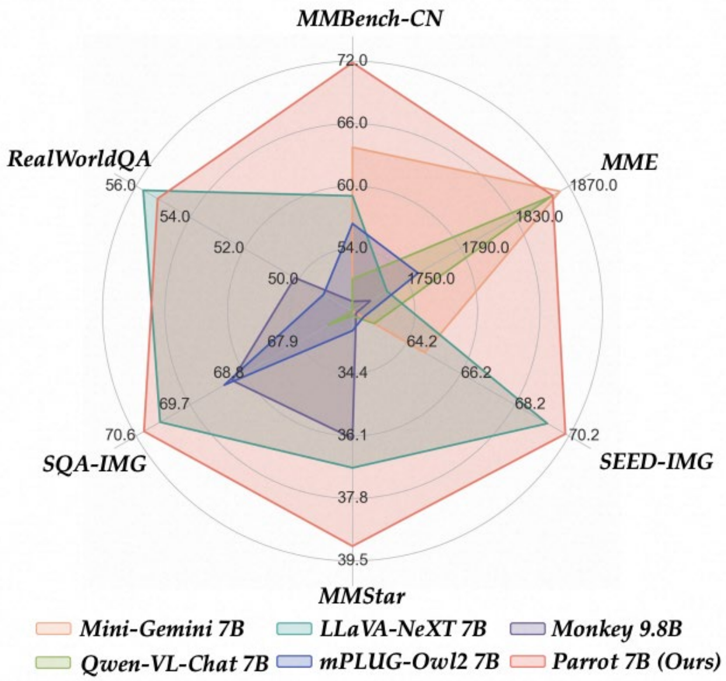
Parrot

基于多语言MOE，将偏向英语的Visual Feature转化为面向特定语言



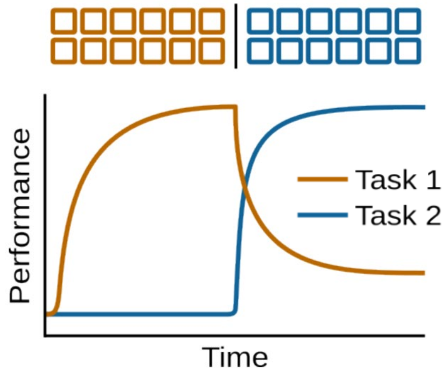
Parrot

Method	LLM	MMMB en	MMMB zh	MMMB pt	MMMB ar	MMMB tr	MMMB ru	MMBench en	MMBench zh	MMBench pt	MMBench ar	MMBench tr	MMBench ru
Open-source models													
LLaVA-1.5 [36]	Vicuna-v1.5-7B	67.07	58.83	59.76	43.50	46.43	59.06	65.37	58.33	59.02	36.16	43.90	56.95
LLaVA-1.5 [36]	Vicuna-v1.5-13B	69.76	62.86	60.76	45.49	54.44	62.69	68.98	63.23	62.97	46.56	53.17	61.59
LLaVA-NeXT [37]	Vicuna-v1.5-7B	70.87	61.57	61.81	42.74	46.95	63.85	67.95	60.56	60.39	38.40	45.36	59.62
LLaVA-NeXT [37]	Vicuna-v1.5-13B	**74.44**	67.19	63.21	45.36	53.09	68.24	70.87	64.51	64.08	45.36	52.92	61.85
Qwen-VL [6]	Qwen-7B	52.63	36.37	38.65	36.54	37.42	40.70	42.26	22.25	25.08	18.72	26.37	28.17
Qwen-VL-Chat [6]	Qwen-7B	56.02	57.77	46.37	43.04	41.05	48.65	54.29	56.52	43.12	35.73	39.17	42.86
MiniGPT-4-v2 [75]	LLaMA2-13B	38.71	30.05	31.52	26.60	26.02	29.23	23.88	11.76	14.26	2.49	6.78	12.54
ShareGPT4V [12]	Vicuna-v1.5-7B	69.24	60.23	60.29	43.57	45.26	61.23	69.59	61.6	59.62	37.37	43.38	59.45
InstructBLIP [17]	Vicuna-7B	39.47	32.92	35.67	23.80	28.36	36.37	27.83	18.81	27.14	3.26	8.50	20.87
mPLUG-Owl2 [64]	LLaMA2-7B	67.25	60.99	59.70	45.78	45.43	62.63	66.15	59.36	58.24	37.88	47.68	60.39
Monkey [35]	Qwen-VL-7B	66.02	58.18	46.31	38.83	37.66	48.59	58.07	53.52	49.57	31.01	31.35	45.18
Monkey-chat [35]	Qwen-VL-7B	71.63	66.54	60.35	48.77	46.31	58.59	70.79	65.72	65.03	46.90	48.10	59.36
VisualGLM [18]	ChatGLM-6B	31.05	18.07	19.42	15.38	22.81	19.77	23.2	17.18	11.43	2.92	6.62	5.33
VisCPM-Chat [24]	CPM-Bee-10B	53.10	47.54	28.19	26.90	26.78	26.84	45.88	46.39	15.81	1.46	9.19	1.20
PARROT	Qwen1.5-7B	70.00	68.13	67.31	62.69	58.01	66.26	70.70	70.36	65.12	57.82	58.43	64.00
PARROT	Qwen1.5-14B	73.92	**71.64**	**69.82**	**68.13**	**64.33**	**70.18**	**74.40**	**72.25**	**69.16**	**66.15**	**64.52**	**69.33**
Closed-source models													
GPT-4V [46]	Private	74.97	74.21	71.46	**73.51**	68.95	73.10	**77.60**	74.40	72.51	72.34	**70.53**	74.83
Gemini Pro [58]	Private	75.03	71.87	70.64	69.94	**69.59**	72.69	73.63	72.08	70.27	61.08	69.76	70.45
Qwen-VL-MAX [6]	Private	**77.19**	**75.26**	**72.16**	70.82	66.02	**74.21**	76.80	**77.58**	**74.57**	**75.00**	69.07	**75.00**



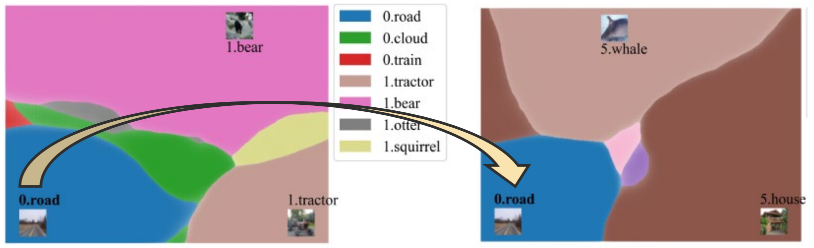
在同量级多语言评测任务上获得领先性能

总结



特征表示兼容

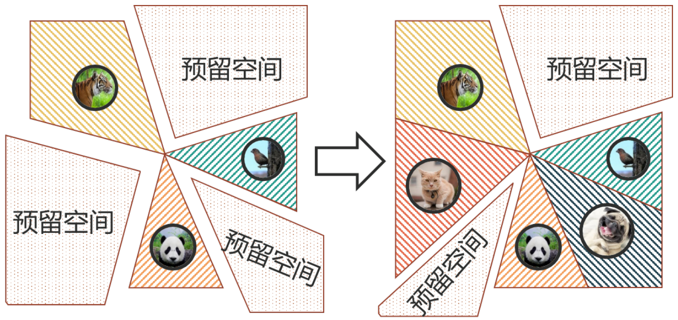
特征表示矫正



(a) Input Space of Model 1

(b) Input Space of Mode

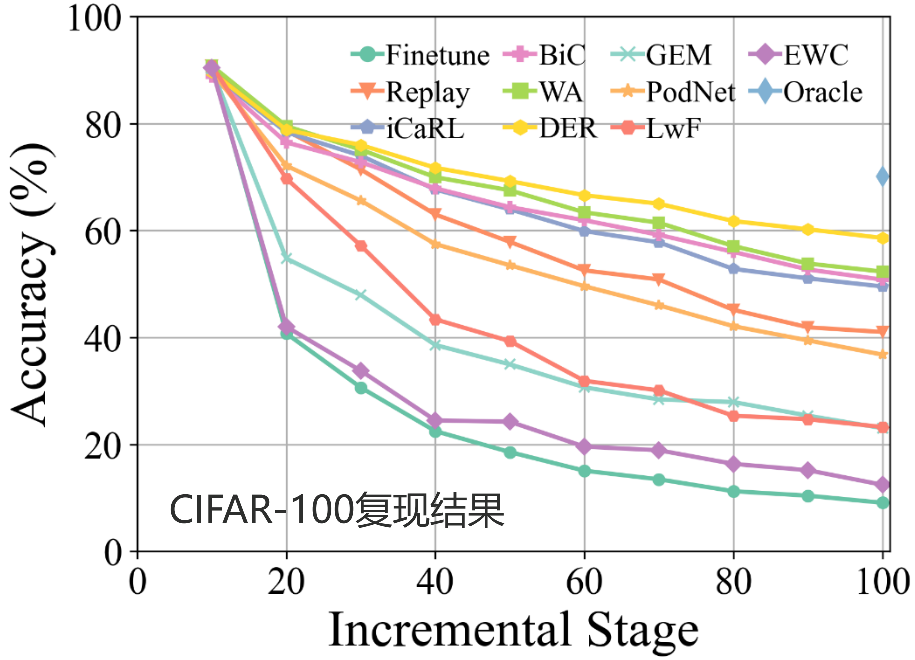
特征表示预留



特征表示扩张

PyCIL:持续学习工具包

https://github.com/LAMDA-CL/PyCIL



全面 · 基准 · 可扩展 · 长期维护

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, De-Chuan Zhan. PyCIL: A Python Toolbox for Class-Incremental Learning. **SCIS** 2022
Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, Ziwei Liu. Class-incremental learning: A survey. **TPAMI** 2024
周大蔚, 汪福运, 叶翰嘉, 詹德川. 基于深度学习的类别增量学习算法综述. **计算机学报** 2023.

PILOT:（基于预训练模型的）持续学习工具包

谢谢

PILOT

A Pre-trained Model-based Continual Learning Toolbox

https://github.com/LAMDA-CL/LAMDA-PILOT

- `FineTune` : Baseline method which simply updates parameters on new tasks.
- `iCaRL` : iCaRL: Incremental Classifier and Representation Learning. CVPR 2017 [paper]
- `Coil` : Co-Transport for Class-Incremental Learning. ACMMM 2021 [paper]
- `DER` : DER: Dynamically Expandable Representation for Class Incremental Learning. CVPR 2021 [paper]
- `FOSTER` : Feature Boosting and Compression for Class-incremental Learning. ECCV 2022 [paper]
- `MEMO` : A Model or 603 Exemplars: Towards Memory-Efficient Class-Incremental Learning. ICLR 2023 Spotlight [paper]
- `SimpleCIL` : Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. arXiv 2023 [paper]
- `L2P` : Learning to Prompt for Continual Learning. CVPR 2022 [paper]
- `DualPrompt` : DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. ECCV 2022 [paper]
- `CODA-Prompt` : CODA-Prompt: COntinual Decomposed Attention-based Prompting for Rehearsal-Free Continual Learning. CVPR 2023 [paper]
- `ADAM` : Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. arXiv 2023 [paper]
- `RanPAC` : RanPAC: Random Projections and Pre-trained Models for Continual Learning. NeurIPS 2023 [paper]
- `Ease` : Expandable Subspace Ensemble for Pre-Trained Model-Based Class-Incremental Learning. CVPR 2024 [paper]

涵盖典型CIL算法与基于预训练模型的最优算法

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan. Continual Learning with Pre-Trained Models: A Survey. **IJCAI** 2024

Hai-Long Sun, Da-Wei Zhou, De-Chuan Zhan, Han-Jia Ye. Pilot: A pre-trained model-based continual learning toolbox. **SCIS** 2025