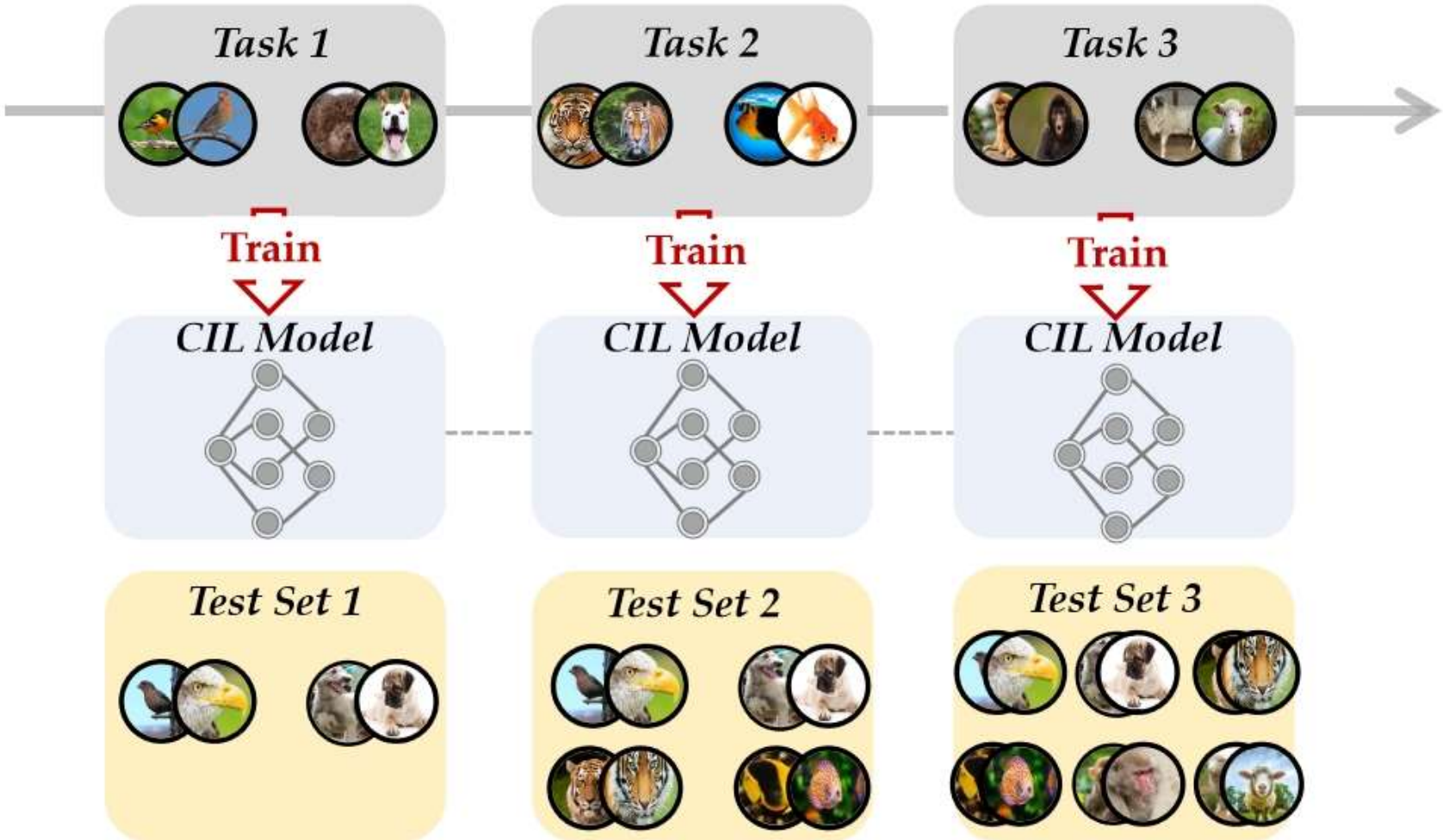
Class-Incremental Learning

Introduction



Catastrophic forgetting

- Data-Centric Class-Incremental Learning
- Model-Centric Class-Incremental Learning
- Algorithm-Centric Class-Incremental Learning

Preliminaries

- There is a sequence of B training tasks

$$\{\mathcal{D}^1, \mathcal{D}^2, \cdots, \mathcal{D}^B\}$$

- b-th incremental step with n training instances

$$\mathcal{D}^b = \left\{\left(\mathbf{x}_i^b, y_i^b\right)\right\}_{i=1}^{n_b}$$

- Every label space haven't overlap

$$Y_b \cap Y_{b'} = \varnothing \; for \; b \neq b'$$

- After each task, the model is evaluated over all seen class.

$$\mathcal{Y}_b = Y_1 \cup \cdots Y_b \qquad f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \mathbb{E}_{(\mathbf{x},y)\sim \mathcal{D}_t^1 \cup \cdots \mathcal{D}_t^b} \mathbb{I}(y \neq f(\mathbf{x}))$$

Exemplars and exemplar set

- Exemplar Set is an extra collection of instances from former tasks

$$\mathcal{E} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{M},\ y_j \in \mathcal{Y}_{b-1}$$

- Exemplar Set Management
  - keep a fixed number of exemplars per class $R|\mathcal{Y}_b|$
  - saving a fixed number of exemplars $[\frac{M}{|\mathcal{Y}_b|}]$
- Exemplar Selection
  - Randomly sample exemplars
  - Herding

$$\mu_y \leftarrow \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)\,. \qquad \|\mu_y - \phi(\mathbf{x}_i)\| \qquad \text{top-}[\frac{M}{|\mathcal{Y}_b|}]$$

Data-Centric Class-Incremental Learning

- Data Replay
- Data Regularization
- Discussions about Data-Centric Methods

Data Replay

- Revisit exemplars
- Exemplar selection
  - Uncertainty
  - Data center
- Exemplars set efficiency
  - Low fidelity feature
- Generative replay
  - The quality of generated data
  - GAN also suffers the catastrophic forgetting

$$\mathcal{L} = \sum_{(\mathbf{x},y)\in(\mathcal{D}^b\cup\mathcal{E})} \ell(f(\mathbf{x}),y)\,.$$

Rainbow Memory: Continual Learning with a Memory of Diverse Samples

Memory Efficient Class-Incremental Learning for Image Classification

Data Regularization

- Optimizing the model for new classes will not hurt former ones
  - GEM

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{(\mathbf{x},y) \in \mathcal{D}^b} \ell\left(f(\mathbf{x}), y\right)$$

$$\text{s.t.} \sum_{(\mathbf{x}_j,y_j) \in \mathcal{E}} \ell\left(f(\mathbf{x}_j), y_j\right) \leq \sum_{(\mathbf{x}_j,y_j) \in \mathcal{E}} \ell\left(f^{b-1}(\mathbf{x}_j), y_j\right)$$

$$\langle g, g_{old} \rangle := \left\langle \frac{\partial \ell\left(f(\mathbf{x}), y\right)}{\partial \theta}, \frac{\partial \ell\left(f, \mathcal{E}\right)}{\partial \theta} \right\rangle \geq 0\,,$$

Discussions about Data-Centric Methods

- Data replay may suffer the overfitting problem
- Data-imbalance problem
- Few-shot exemplars VS. Many-shot task
- How to select data as exemplar

Algorithm 1 Diversity-Aware Memory Update

Input: K denotes memory size, N_t denotes the number of seen classes until task t, $\mathfrak{D}_t^S$ denotes stream data at task t, $\mathfrak{D}_{t-1}^M$ denotes exemplars stored in a episodic memory after task $t-1$.
Output: $\mathfrak{D}_t^M$ exemplars after learning task t.
$\mathfrak{D}_t^M = \{\}$ ▷ New exemplars from scratch
$k_c = floor(K/N_t)$ ▷ Class-balanced sampling
for $c = 1, 2, \ldots, N_t$ **do**
    $\mathfrak{D}_c = \{(x,y) | y = c, (x,y) \in \mathfrak{D}_t^S \cup \mathfrak{D}_{t-1}^M\}$
    Sort $\mathfrak{D}_c$ by $u(x)$ computed by (4)
    for $j = 1, 2, \ldots, k_c$ **do**
        $i = j * |\mathfrak{D}_c| / k_c$ ▷ $|\mathfrak{D}_c|/k_c$ step-size indexing
        $\mathfrak{D}_t^M$ += $\mathfrak{D}_c[i]$
    end for
end for

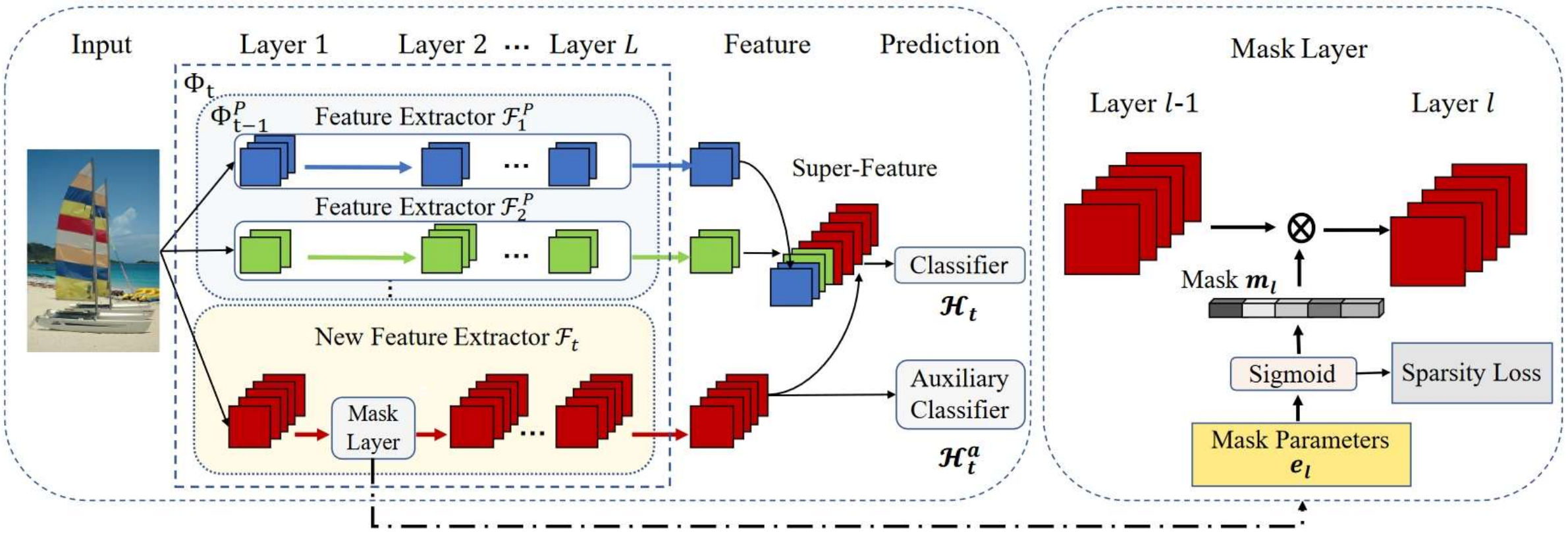
Model-Centric Class-Incremental Learning

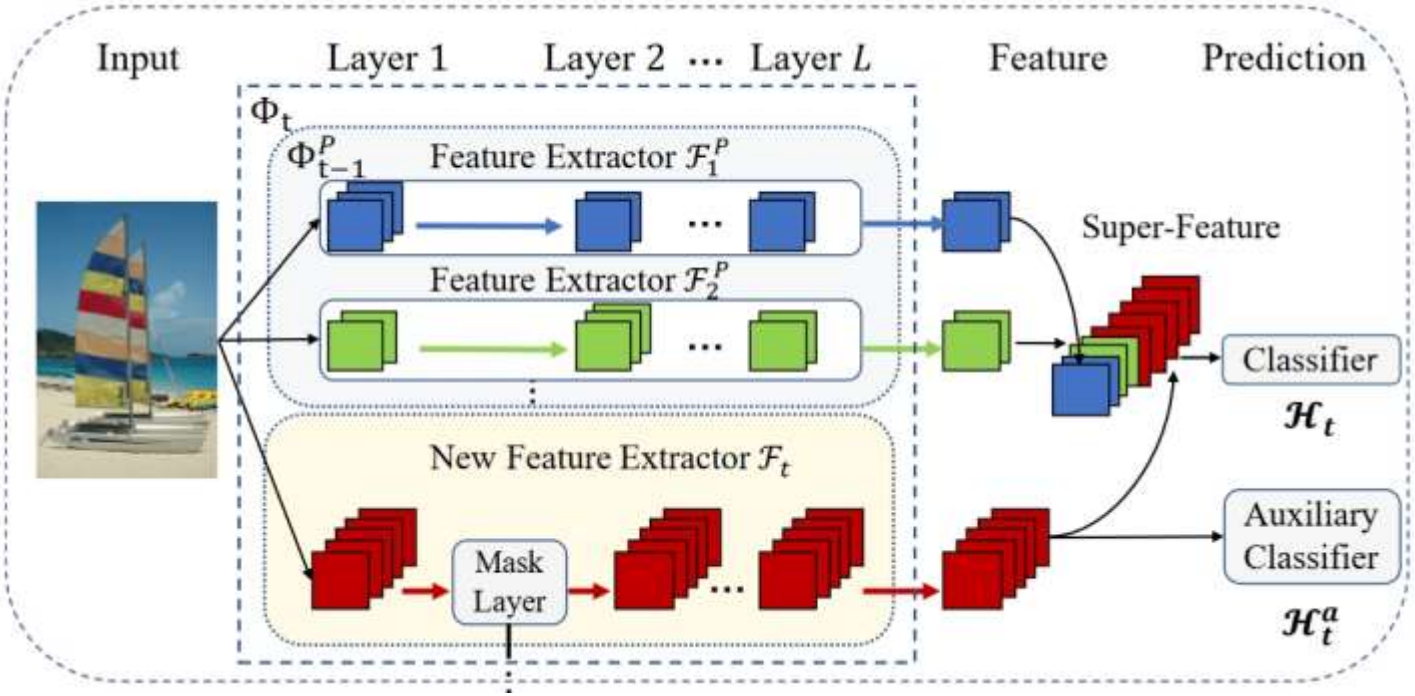
- Dynamic Networks
- Parameter Regularization

Dynamic Networks

- Expands a new backbone when facing new tasks and aggregates the features with a larger FC layer DER

DER


Input
Layer 1
Layer 2 ··· Layer L
Feature
Prediction
Φ_t
Φ_{t-1}^P
Feature Extractor $\mathcal{F}_1^P$
Feature Extractor $\mathcal{F}_2^P$
New Feature Extractor $\mathcal{F}_t$
Mask Layer
Super-Feature
Classifier
$\mathcal{H}_t$
Auxiliary Classifier
$\mathcal{H}_t^a$
Mask Layer
Layer l-1
Layer l
Mask m_l
Sigmoid
Sparsity Loss
Mask Parameters e_l



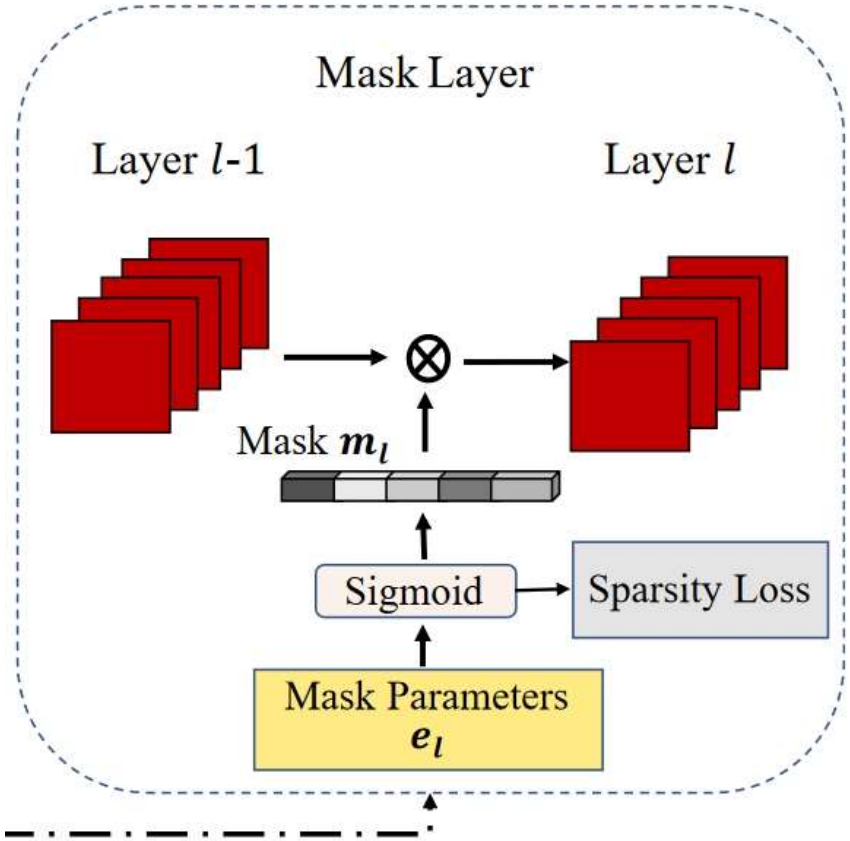
$$\mathcal{L}_{\mathcal{H}_t} = -\frac{1}{|\tilde{\mathcal{D}}_t|}\sum_{i=1}^{|\tilde{\mathcal{D}}_t|}\log(p_{\mathcal{H}_t}(y=y_i|\boldsymbol{x_i}))$$

$$|\mathcal{Y}_t|+1$$

$$\mathcal{L}_{\mathrm{ER}} = \mathcal{L}_{\mathcal{H}_t} + \lambda_a \mathcal{L}_{\mathcal{H}_t^a}$$

$$\boldsymbol{u} = \Phi_t(\boldsymbol{x}) = [\Phi_{t-1}(\boldsymbol{x}), \mathcal{F}_t(\boldsymbol{x})]$$

$$p_{\mathcal{H}_t}(\boldsymbol{y}|\boldsymbol{x}) = \mathrm{Softmax}(\mathcal{H}_t(\boldsymbol{u}))$$



$$\tilde{\boldsymbol{u}} = \Phi_t^P(\boldsymbol{x}) = [\mathcal{F}_1^P(\boldsymbol{x}), \mathcal{F}_2^P(\boldsymbol{x}), ..., \phi_t(\boldsymbol{x})]$$

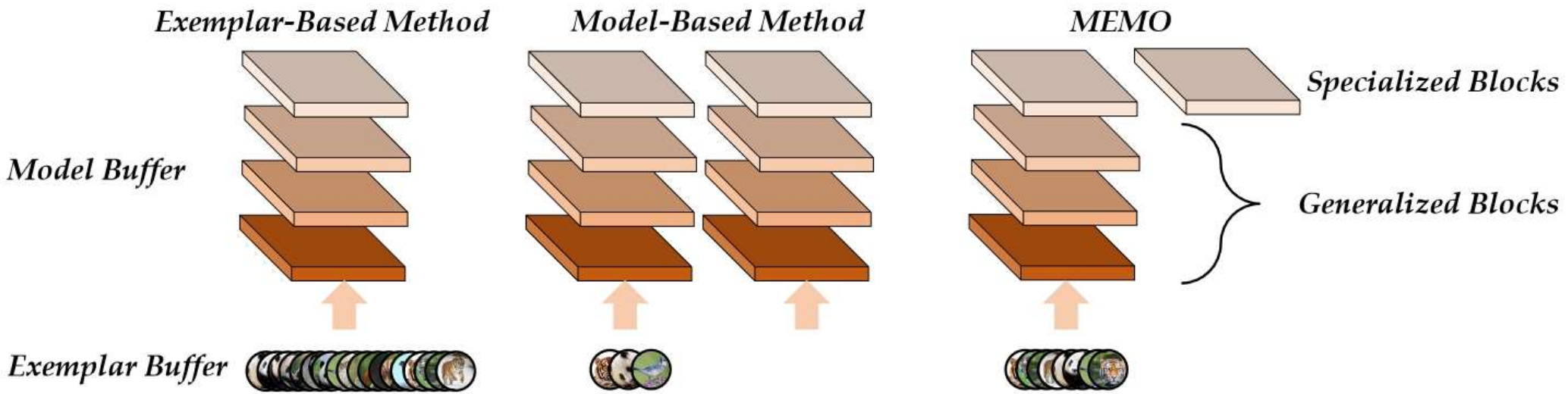
$$\boldsymbol{m_l} = \sigma(s\boldsymbol{e_l})$$

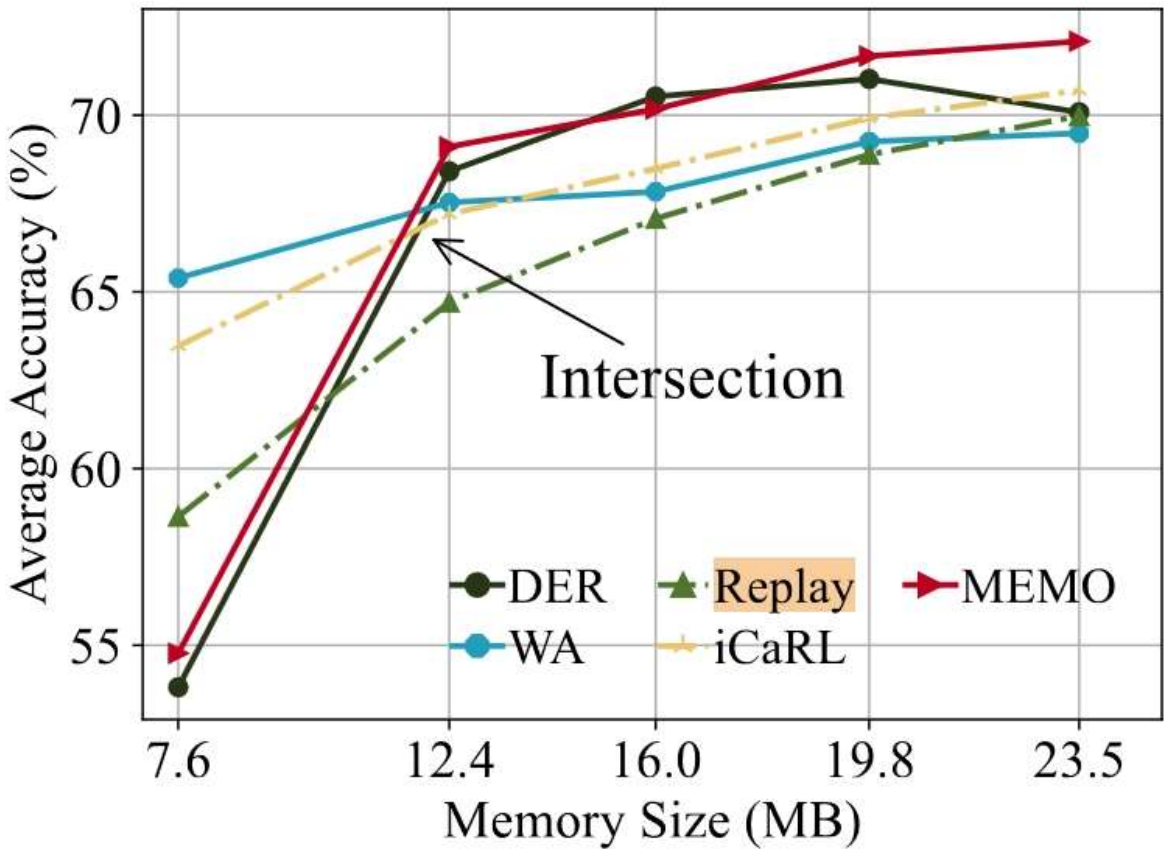
$$s = \frac{1}{s_{\max}} + (s_{\max} - \frac{1}{s_{\max}})\frac{b-1}{B-1}$$

$$\mathcal{L}_S = \frac{\sum_{l=1}^{L} K_l \|\boldsymbol{m_{l-1}}\|_1 \|\boldsymbol{m_l}\|_1}{\sum_{l=1}^{L} K_l c_{l-1} c_l}$$

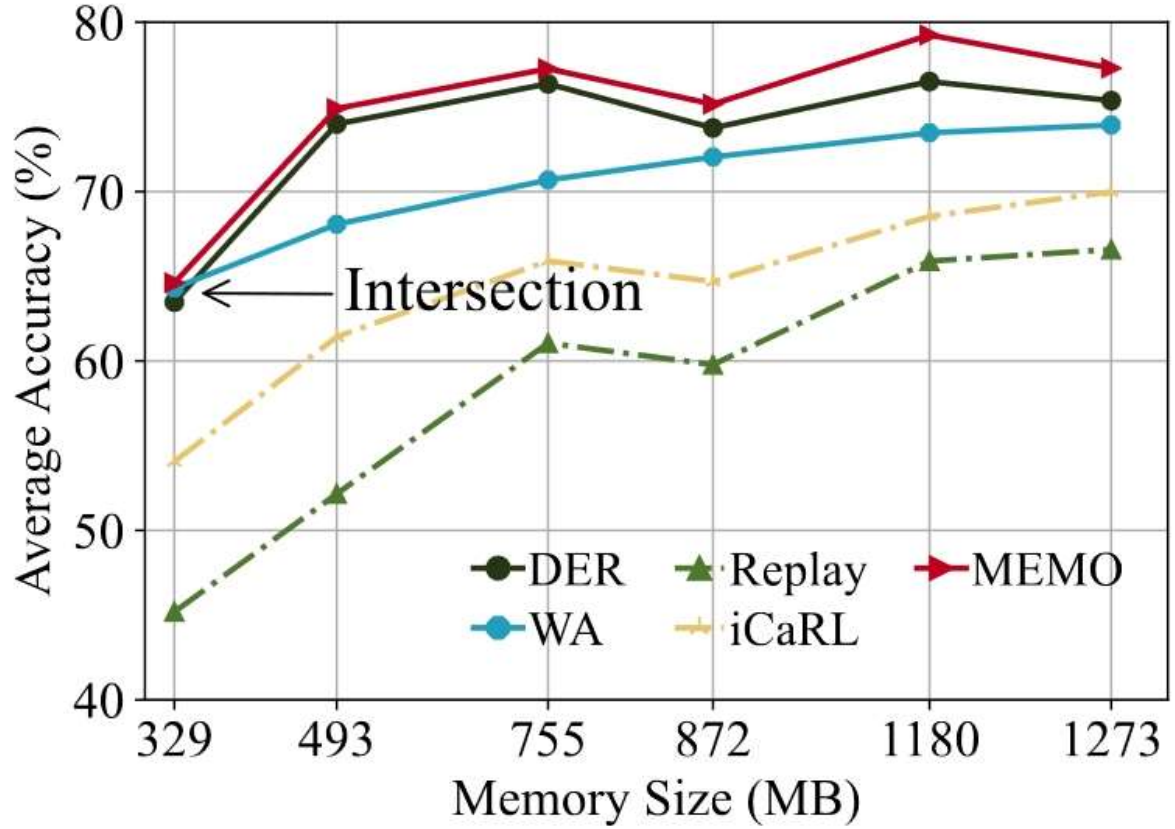
$$\mathcal{L}_{\text{DER}} = \mathcal{L}_{\mathcal{H}_t} + \lambda_a \mathcal{L}_{\mathcal{H}_t^a} + \lambda_s \mathcal{L}_S$$

MEMO

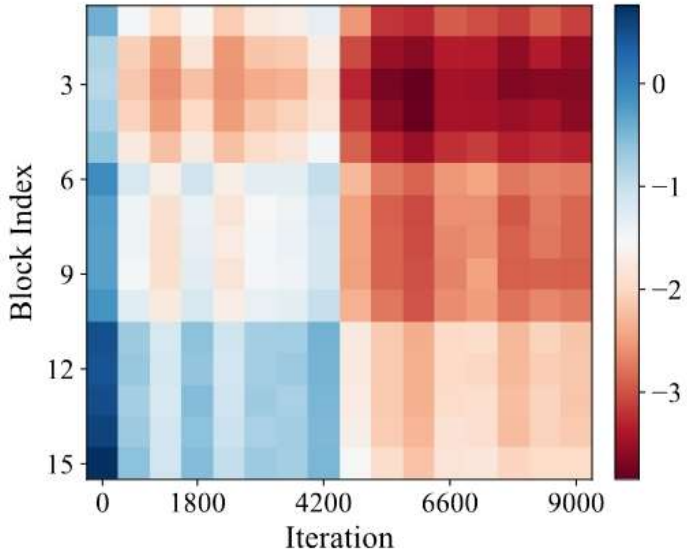

Exemplar-Based Method
Model-Based Method
MEMO
Specialized Blocks
Generalized Blocks
Model Buffer
Exemplar Buffer



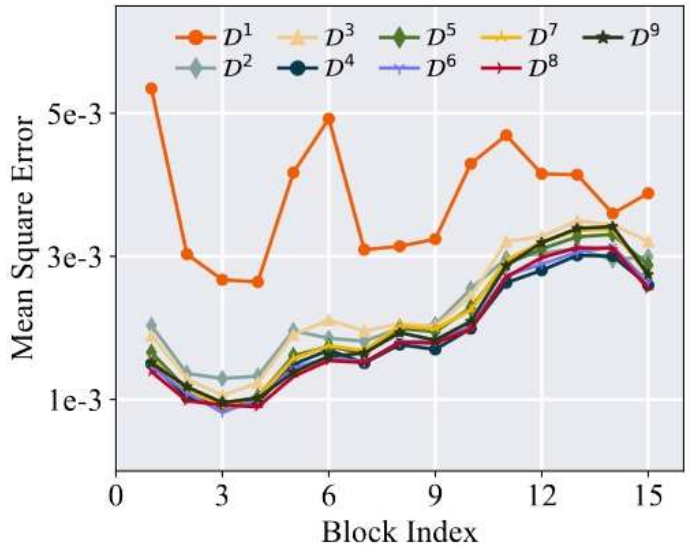
(a) CIFAR100, Base0 Inc10



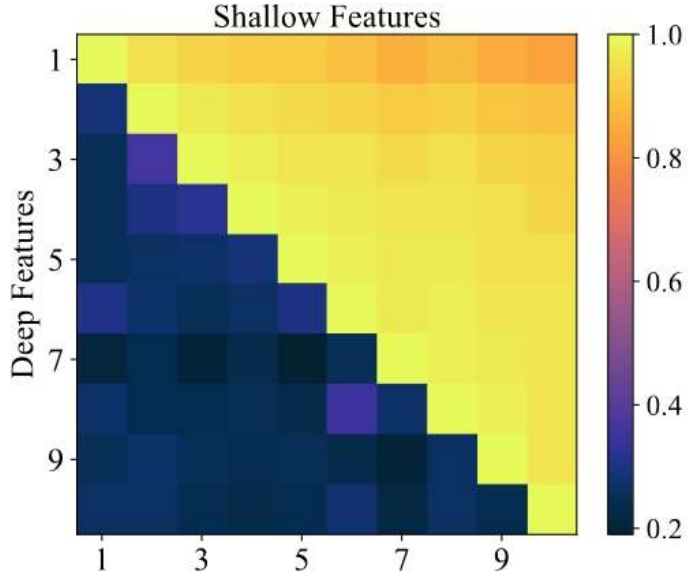
(b) ImageNet100, Base50 Inc5



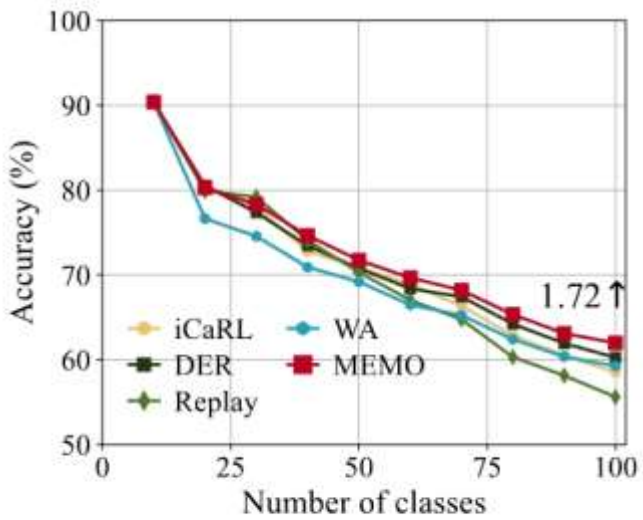
(a) Gradient norm (log scale)



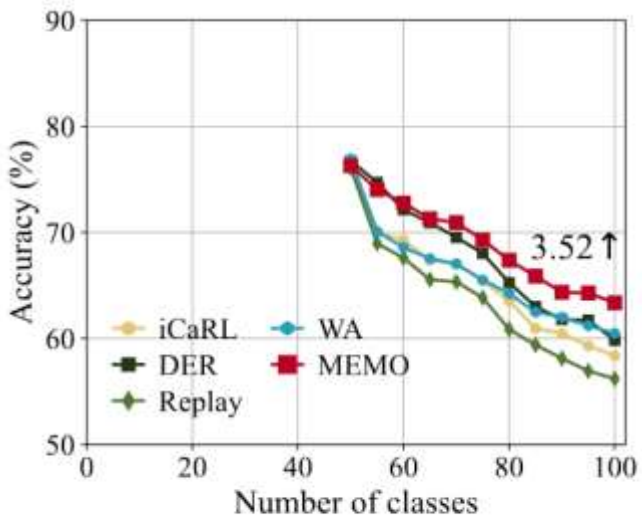
(b) MSE of different blocks



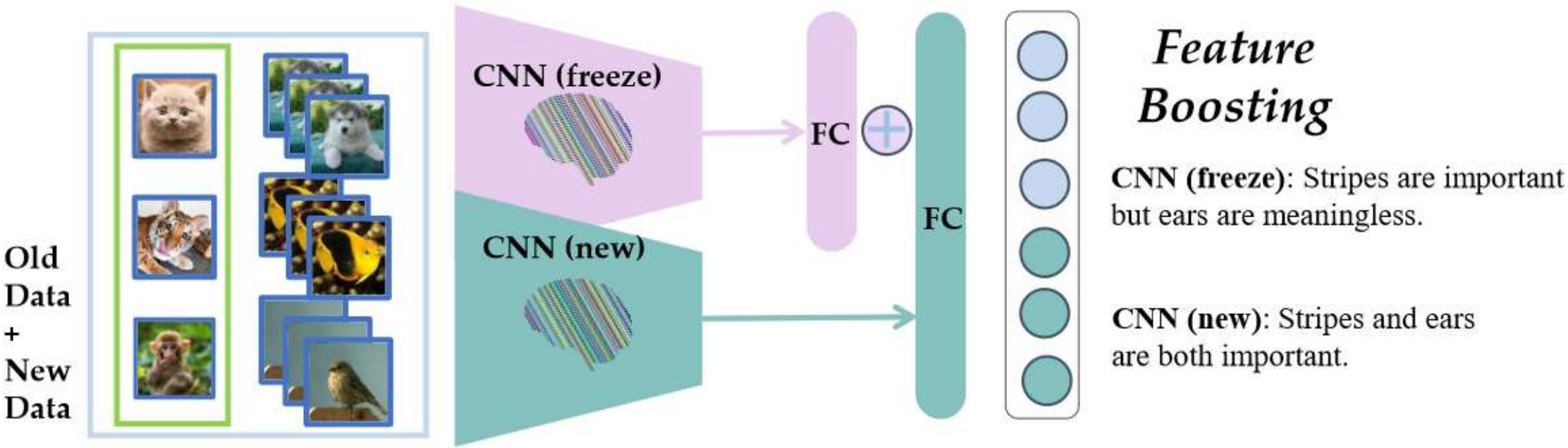
(c) CKA between backbones



(b) CIFAR100 Base0 Inc10



(c) CIFAR100 Base50 Inc5



Foster: Feature Boosting and Compression for Class-Incremental Learning

Parameter Regularization

- keep the important parameters static to maintain former knowledge

$$\mathcal{L} = \ell(f(\mathbf{x}), y) + \frac{1}{2}\lambda \sum_k \Omega_k (\theta_k^{b-1} - \theta_k)^2 .$$

  - Fisher matrix
  - Prompt

Algorithm-Centric Class-Incremental Learning

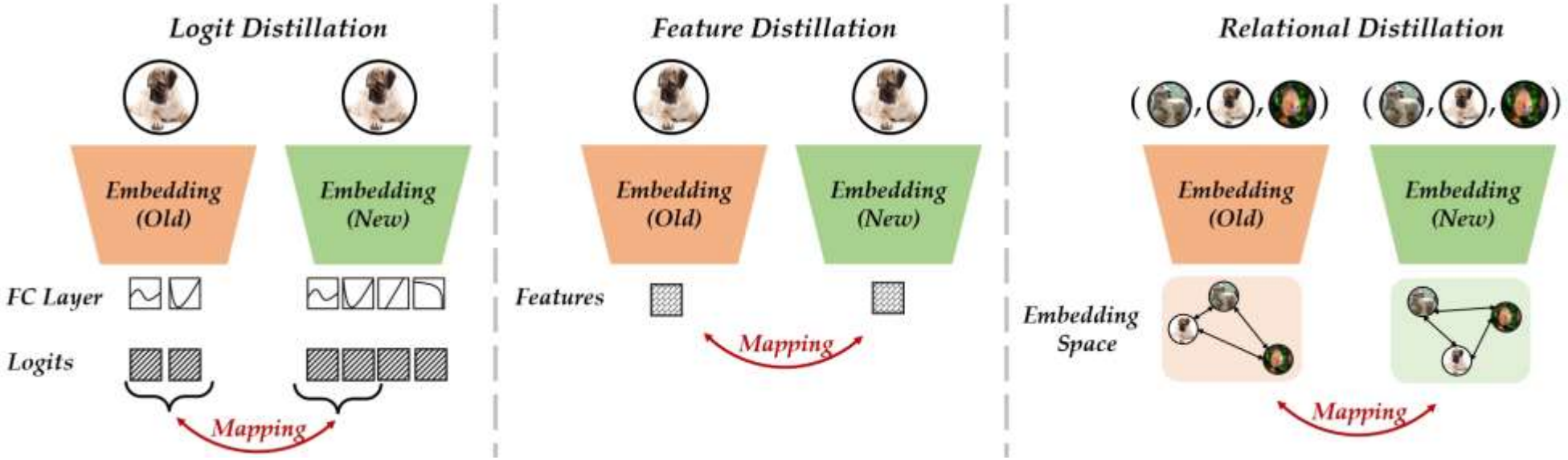
- Knowledge Distillation
- Model Rectify

Knowledge Distillation

$$\mathcal{L} = \ell(f(\mathbf{x}), y) + (1 - \langle \frac{\phi^{b-1}(\mathbf{x})}{\|\phi^{b-1}(\mathbf{x})\|}, \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|} \rangle) .$$

$$\mathcal{L} = \underbrace{\ell(f(\mathbf{x}), y)}_{\text{Learning New Classes}} + \underbrace{\sum_{k=1}^{|\mathcal{Y}_{b-1}|} -\mathcal{S}_k(f^{b-1}(\mathbf{x})) \log \mathcal{S}_k(f(\mathbf{x}))}_{\text{Remembering Old Classes}} , \quad (14)$$

$$\sum_{\{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k\} \in \mathcal{D}^b} \|\cos \angle \mathbf{t}_i \mathbf{t}_j \mathbf{t}_k - \cos \angle \mathbf{s}_i \mathbf{s}_j \mathbf{s}_k\| ,$$



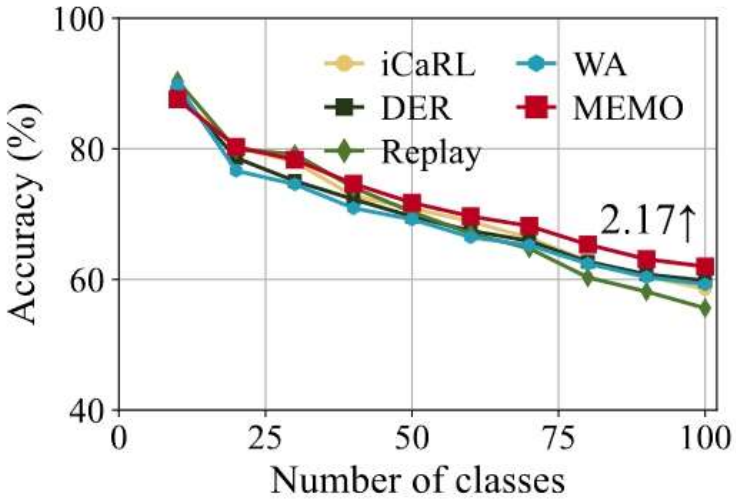
Model Rectify

- Try to find the abnormal behaviors in CIL models and rectify them like the oracle model
  - the weight norm of new classes is significantly larger than old ones
  - the model tends to predict instances as the new classes with larger weights
  - weight drifting
  - logits of new classes are much larger than old ones

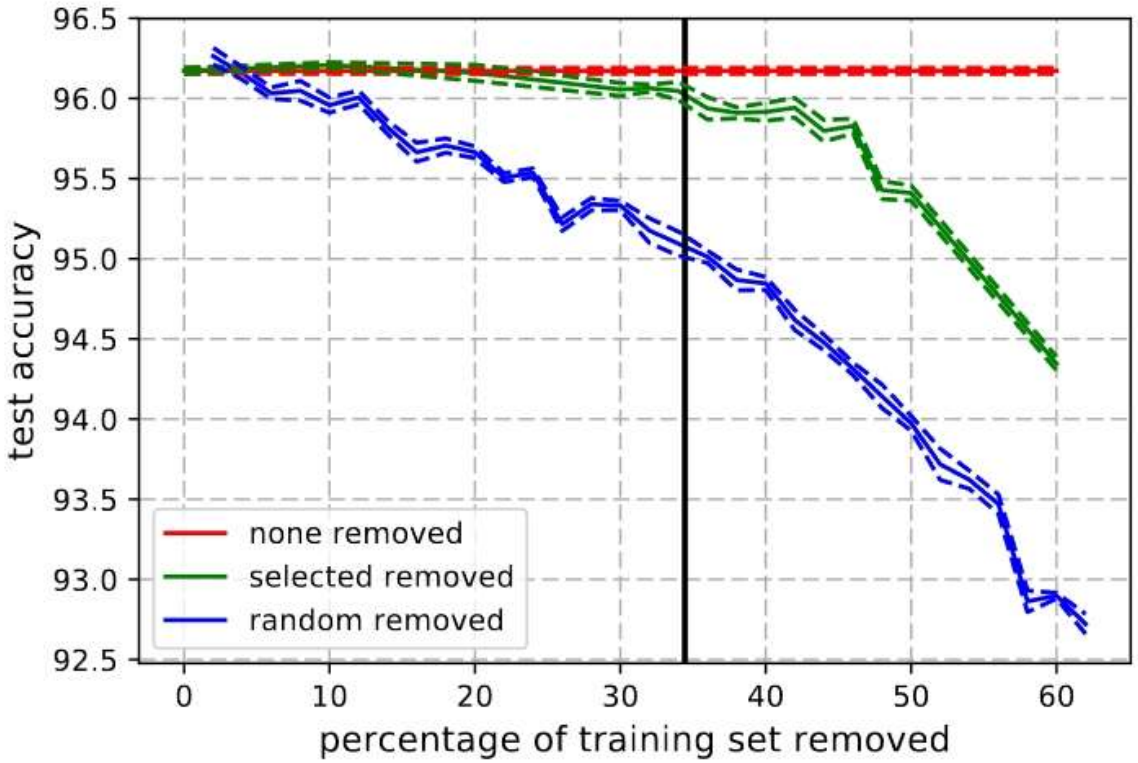
- Data imbalance
- How to select exemplar

Methods	5 steps		10 steps		20 steps		50 steps	
	#Paras	Avg	#Paras	Avg	#Paras	Avg	#Paras	Avg
Bound	11.2	80.40	11.2	80.41	11.2	81.49	11.2	81.74
iCaRL[27]	11.2	$71.14_{\pm 0.34}$	11.2	$65.27_{\pm 1.02}$	11.2	$61.20_{\pm 0.83}$	11.2	$56.08_{\pm 0.83}$
UCIR[12]	11.2	$62.77_{\pm 0.82}$	11.2	$58.66_{\pm 0.71}$	11.2	$58.17_{\pm 0.30}$	11.2	$56.86_{\pm 3.74}$
BiC[12]	11.2	$73.10_{\pm 0.55}$	11.2	$68.80_{\pm 1.20}$	11.2	$66.48_{\pm 0.32}$	11.2	$62.09_{\pm 0.85}$
WA[39]	11.2	$72.81_{\pm 0.28}$	11.2	$69.46_{\pm 0.29}$	11.2	$67.33_{\pm 0.15}$	11.2	$64.32_{\pm 0.28}$
PODNet[6]	11.2	$66.70_{\pm 0.64}$	11.2	$58.03_{\pm 1.27}$	11.2	$53.97_{\pm 0.85}$	11.2	$51.19_{\pm 1.02}$
RPSNet[26]	60.6	70.5	56.5	68.6	-	-	-	-
Ours(w/o P)	33.6	**76.80**$_{\pm 0.79}$(+3.7)	61.6	**75.36**$_{\pm 0.36}$(+5.9)	117.6	**74.09**$_{\pm 0.33}$(+6.76)	285.6	**72.41**$_{\pm 0.36}$(+8.09)
Ours	**2.89**	**75.55**$_{\pm 0.65}$(+2.45)	**4.96**	**74.64**$_{\pm 0.28}$(+5.18)	**7.21**	**73.98**$_{\pm 0.36}$(+6.65)	**10.15**	**72.05**$_{\pm 0.55}$(+7.73)

23.5MB	$\lvert\mathcal{E}\rvert$	$S(\mathcal{E})$	Model Type	# Parameters	Model Size
Replay	7400	21.67MB	ResNet32	0.46M	1.76MB
iCaRL	7400	21.67MB	ResNet32	0.46M	1.76MB
WA	7400	21.67MB	ResNet32	0.46M	1.76MB
DER	2000	5.85MB	ResNet32	4.60M	17.6MB
MEMO	3300	9.66MB	ResNet32	3.62M	13.83MB

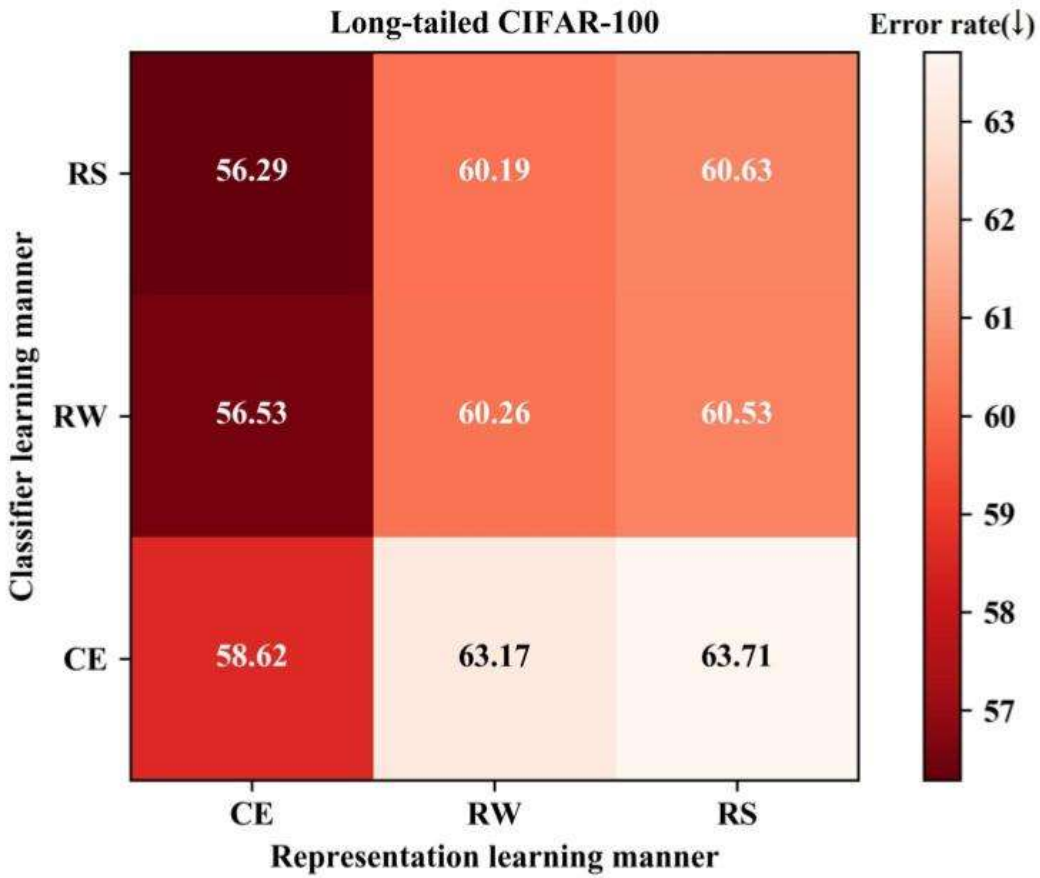


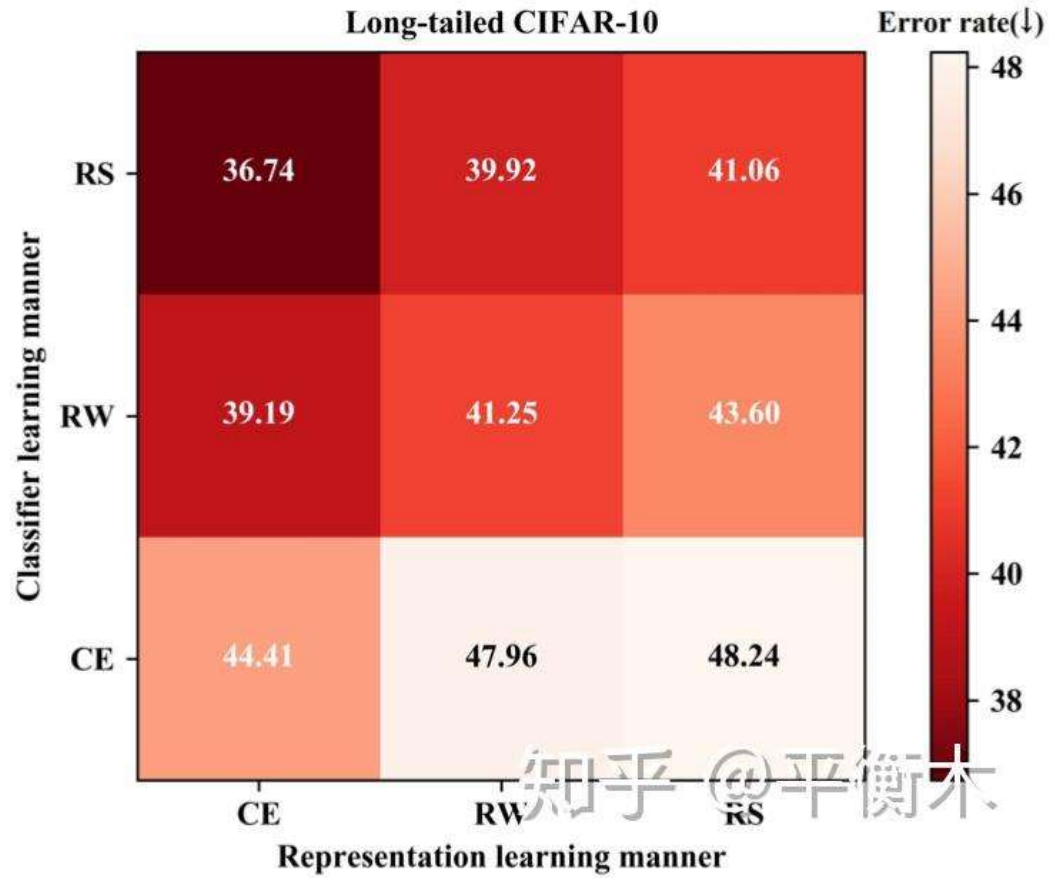
AN EMPIRICAL STUDY OF EXAMPLE FORGETTING DURING DEEP NEURAL NETWORK LEARNING



Algorithm 1 Computing forgetting statistics.

initialize $\text{prev\_acc}_i = 0, i \in \mathcal{D}$
initialize forgetting $T[i] = 0, i \in \mathcal{D}$
while not training done **do**
  $B \sim \mathcal{D}$ # sample a minibatch
  for example $i \in B$ **do**
    compute acc_i
    if $\text{prev\_acc}_i > \text{acc}_i$ **then**
      $T[i] = T[i] + 1$
    $\text{prev\_acc}_i = \text{acc}_i$
  gradient update classifier on B
return T


Long-tailed CIFAR-100
Error rate(↓)
RS
RW
CE
56.29
60.19
60.63
56.53
60.26
60.53
58.62
63.17
63.71
63
62
61
60
59
58
57
Classifier learning manner
CE
RW
RS
Representation learning manner


Long-tailed CIFAR-10
Error rate(↓)
RS
RW
CE
36.74
39.92
41.06
39.19
41.25
43.60
44.41
47.96
48.24
48
46
44
42
40
38
Classifier learning manner
CE
RW
RS
Representation learning manner