终身机器学习

Lifelong Machine Learning

3 类增量学习

主讲：梁国强
gqliang@nwpu.edu.cn

Outlines

3.2 基于正则化的类增量学习

- 基于回放的CL

  旧任务数据不可见→保存部分样本、生成旧任务数据

- 基于正则化的CL

  无法计算旧任务风险→添加损失约束，保留旧任务知识

- 基于网络结构的CL

  模型能力弱→ 扩展网络结构，每一个任务是一个子网络

3.2 基于正则化的类增量学习

■ 基于正则化的CL

无法计算旧任务风险→添加损失约束，保留旧任务知识

目标 $$\sum_{t=1}^{\mathcal{T}} \mathop{\mathbb{E}}_{(\mathcal{X}^{(t)},\mathcal{Y}^{(t)})} [\mathcal{L}(f_t(\mathcal{X}^{(t)};\theta),\mathcal{Y}^{(t)})$$

实际优化 $$\frac{1}{N_{\mathcal{T}}}\sum_{i=1}^{N_{\mathcal{T}}} \ell(f(x_i^{(\mathcal{T})};\theta),y_i^{(\mathcal{T})})$$ +penalty Term

在新任务学习时，保留知识

Distillation

先验模型（参数重要性）

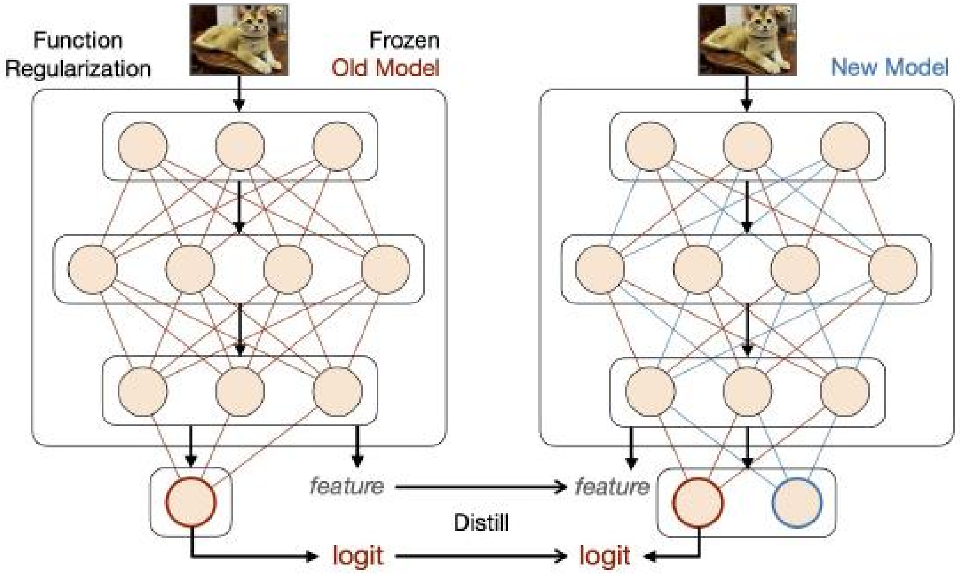
3.2 基于正则化的类增量学习

■ 基于正则化的CL

实际优化 $\frac{1}{N_{\mathcal{T}}}\sum_{i=1}^{N_{\mathcal{T}}}\ell(f(x_i^{(\mathcal{T})};\theta),y_i^{(\mathcal{T})})$ +penalty Term

Distillation

先验模型（参数重要性）



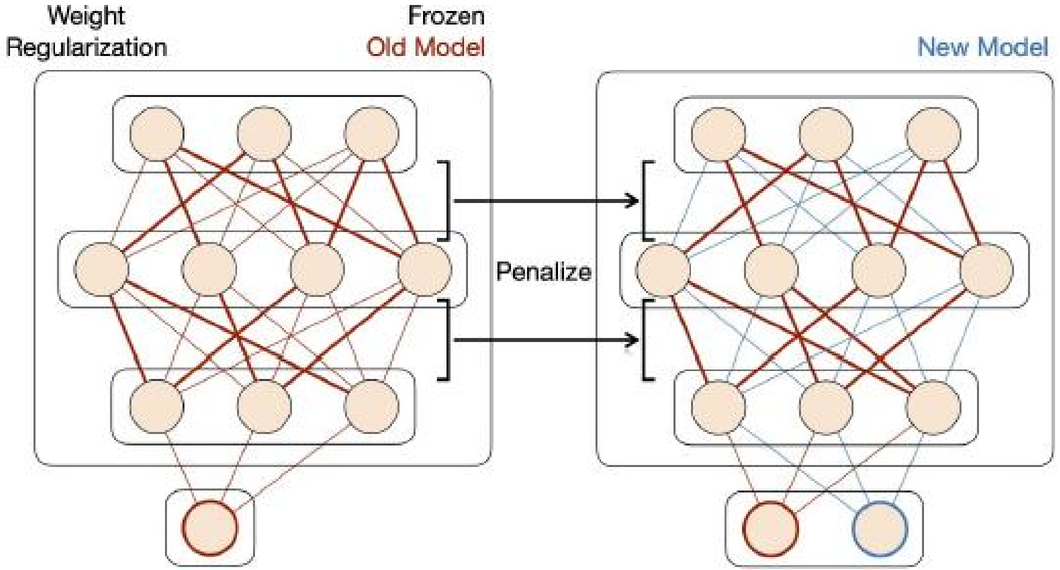
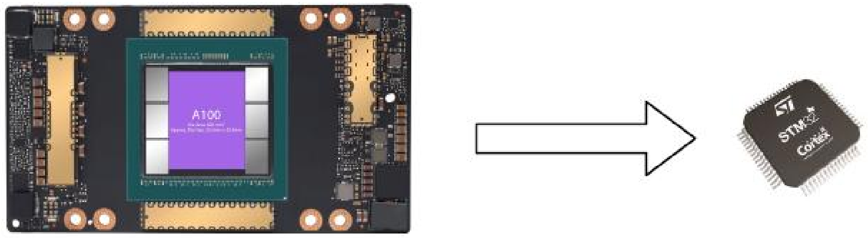


Figure From “A Comprehensive Survey of Continual Learning: Theory, Method and Application”

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation



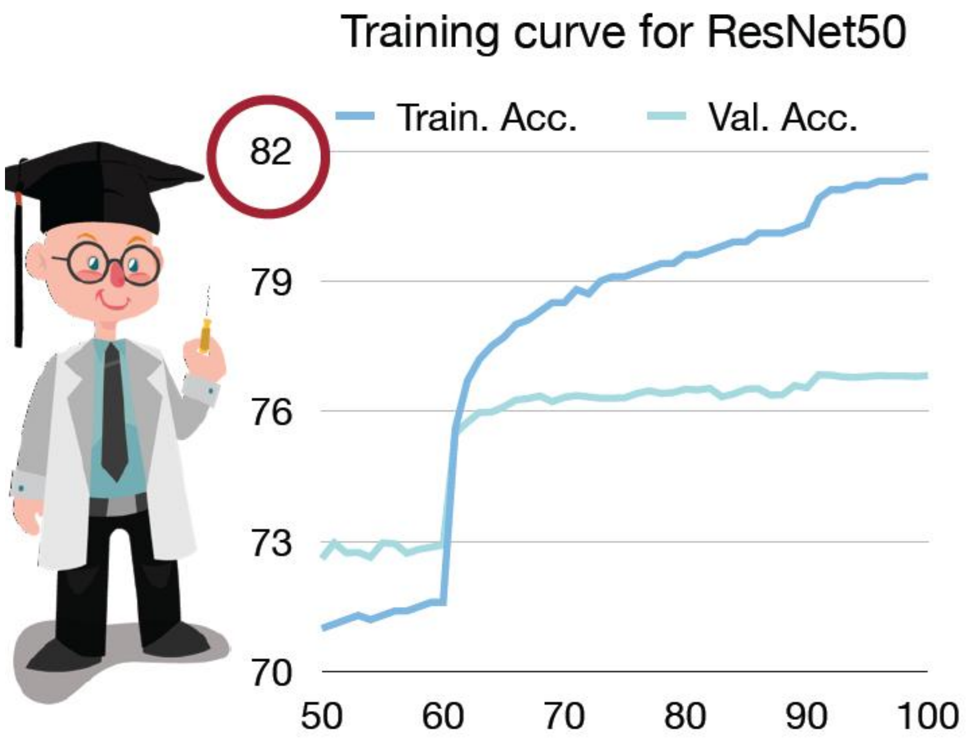
	Cloud AI	Tiny AI
Computation (fp32)	19.5 TFLOPS	MFLOPs
Memory	80GB	256kB
Neural Network	ResNet ViT-Large	MCUNet MobileNetV2-Tiny

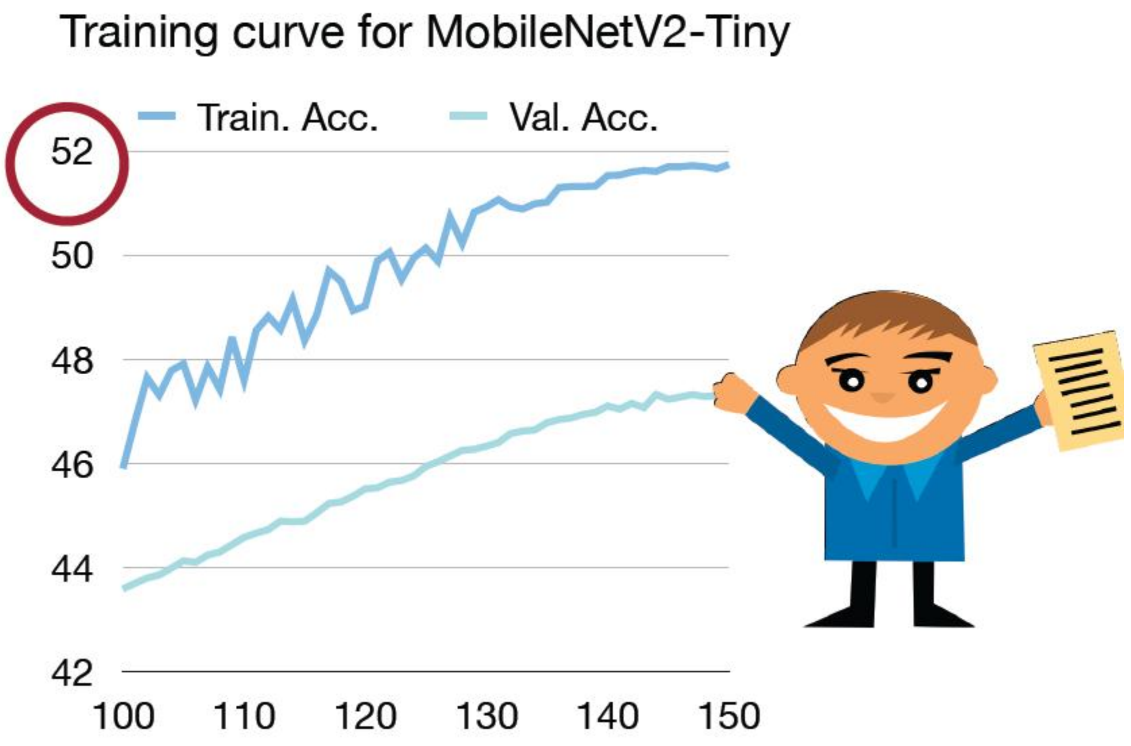
3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation

能不能利用大模型训练小模型

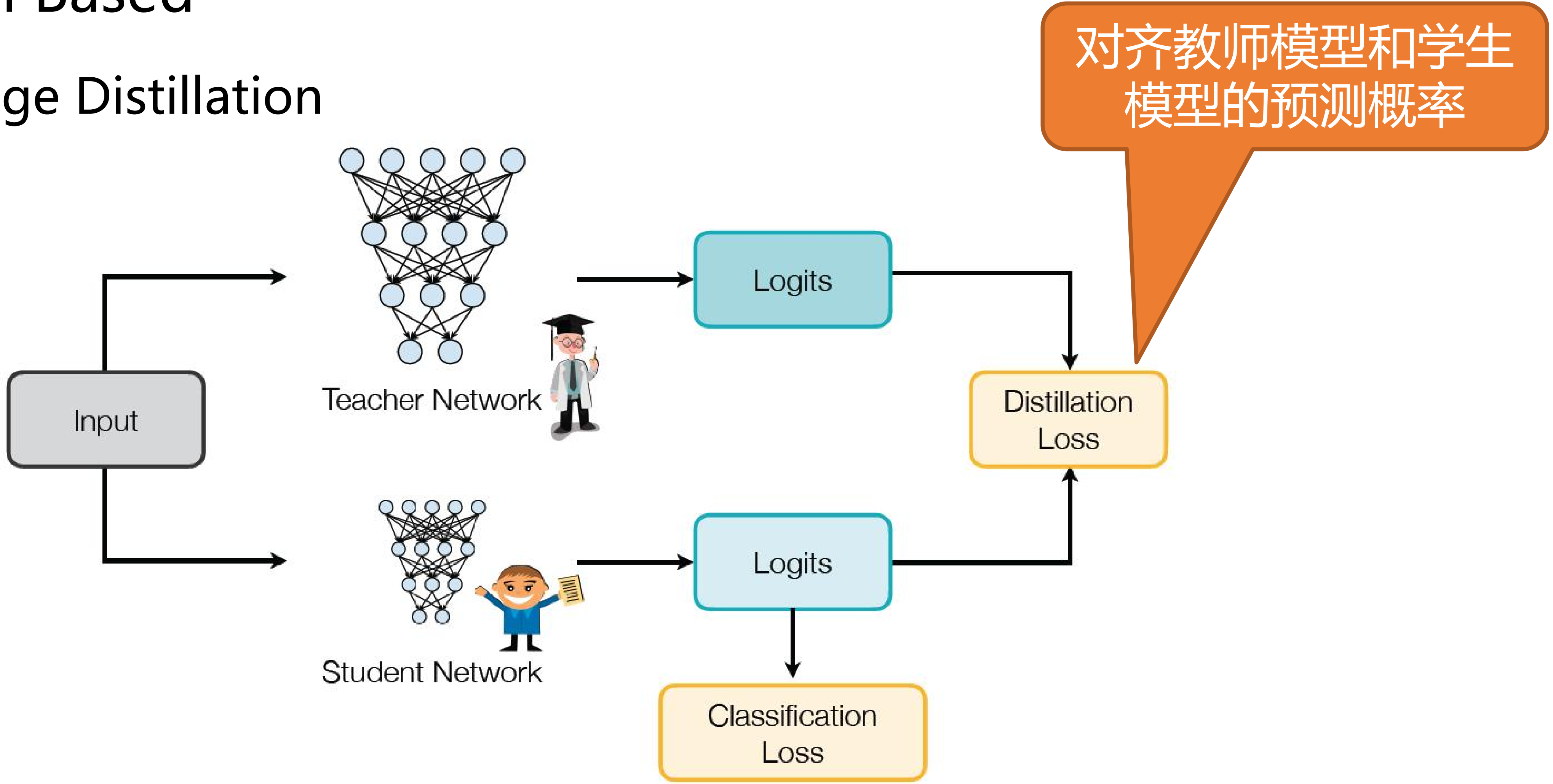




3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation

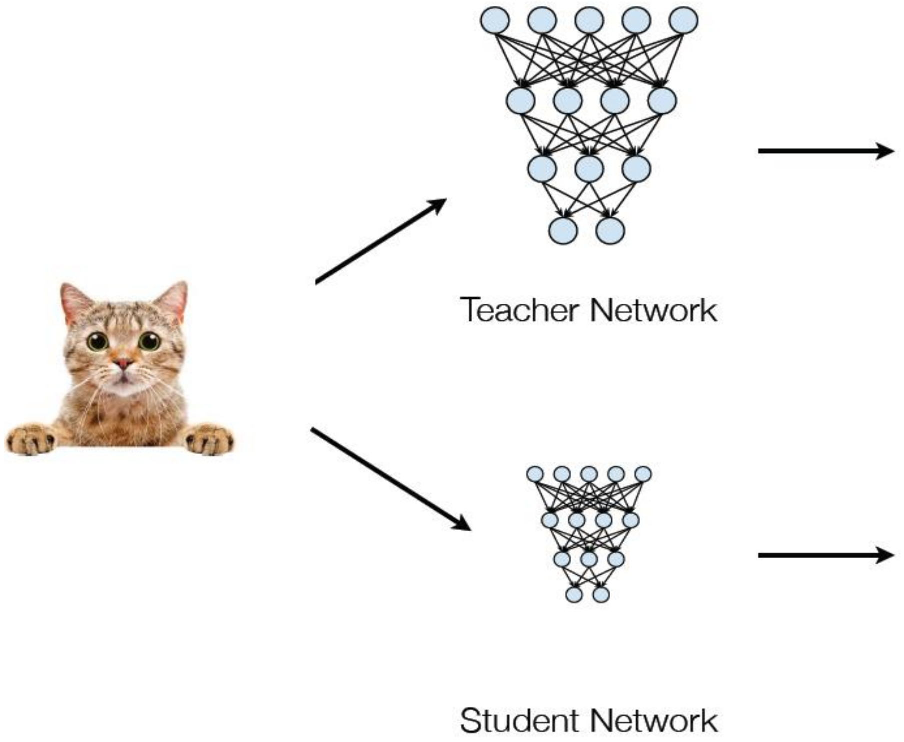


Hinton, Geoffrey et al "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation



	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$\frac{\exp(5)}{\exp(5)+\exp(1)}$$

$$\frac{\exp(1)}{\exp(5)+\exp(1)}$$

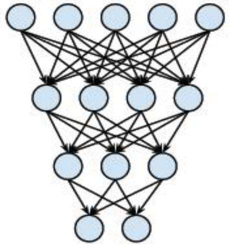
	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation

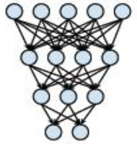


Teacher Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$\frac{\exp(5)}{\exp(5)+\exp(1)}$$

$$\frac{\exp(1)}{\exp(5)+\exp(1)}$$



Student Network

	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z:预测logit T: 温度系数

Distillation Based

- Knowledge Distillation

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z:预测logit T: 温度系数

温度系数用于平滑预测分布，T越大分布越平滑

$$\frac{\exp(5/1)}{\exp(5/1)+\exp(1/1)}$$

	Logits	Probabilities (T=1)	Probabilities (T=10)
Cat	5	0.982	0.599
Dog	1	0.017	0.401

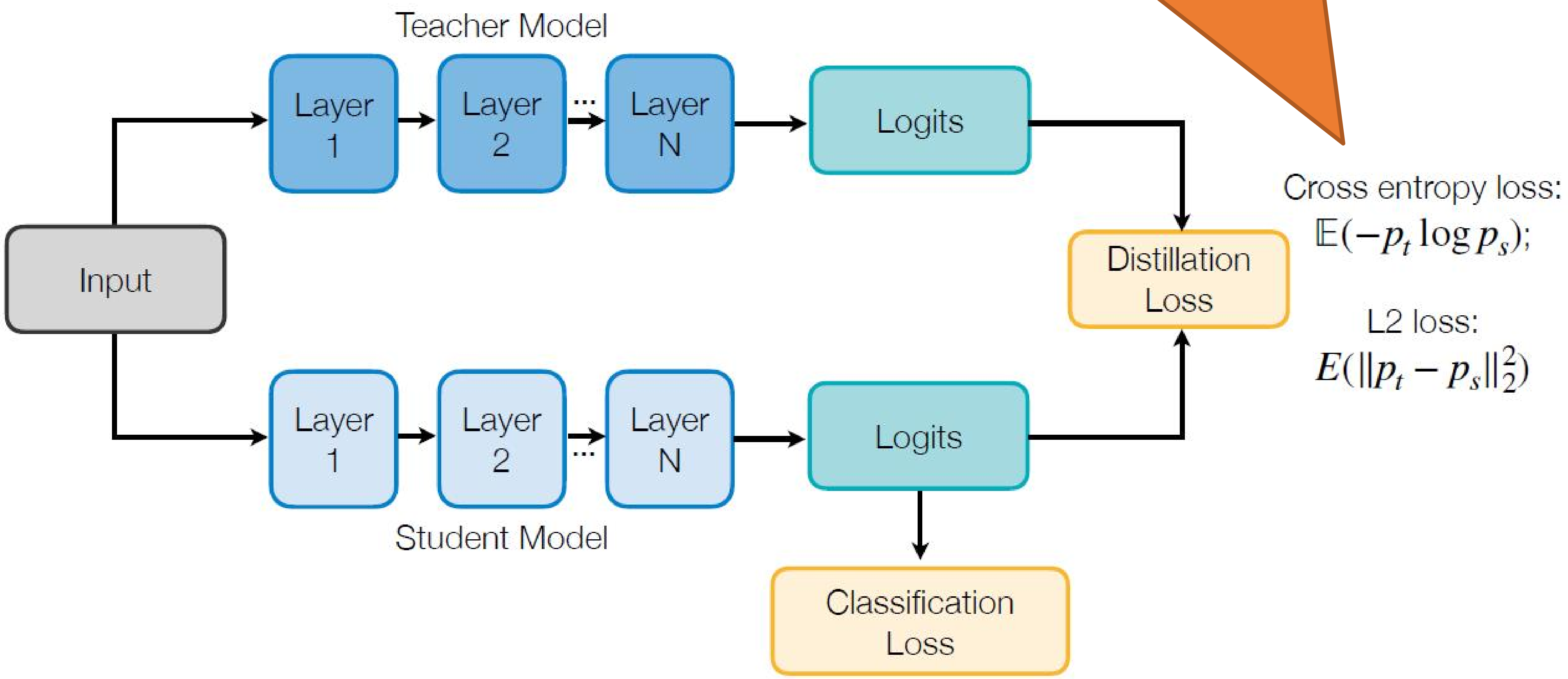
$$\frac{\exp(5/10)}{\exp(5/10)+\exp(1/10)}$$

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation——对齐Logits

Soft-Target降低学习难度
挖掘Non-target的信息

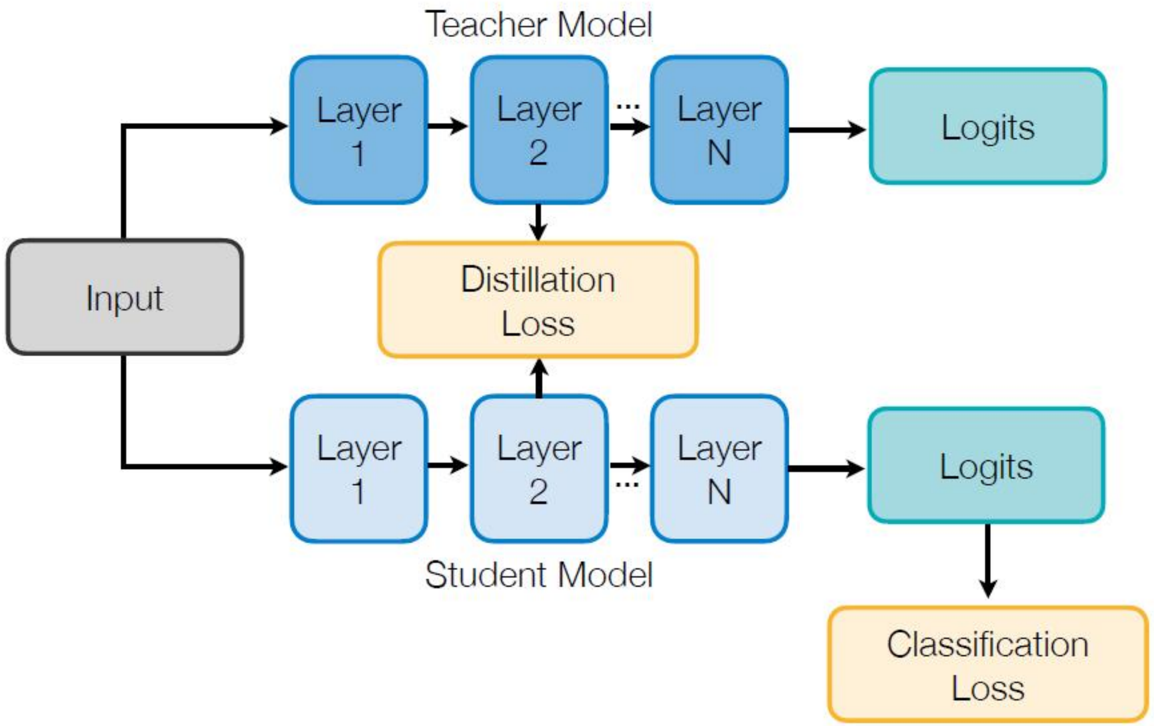


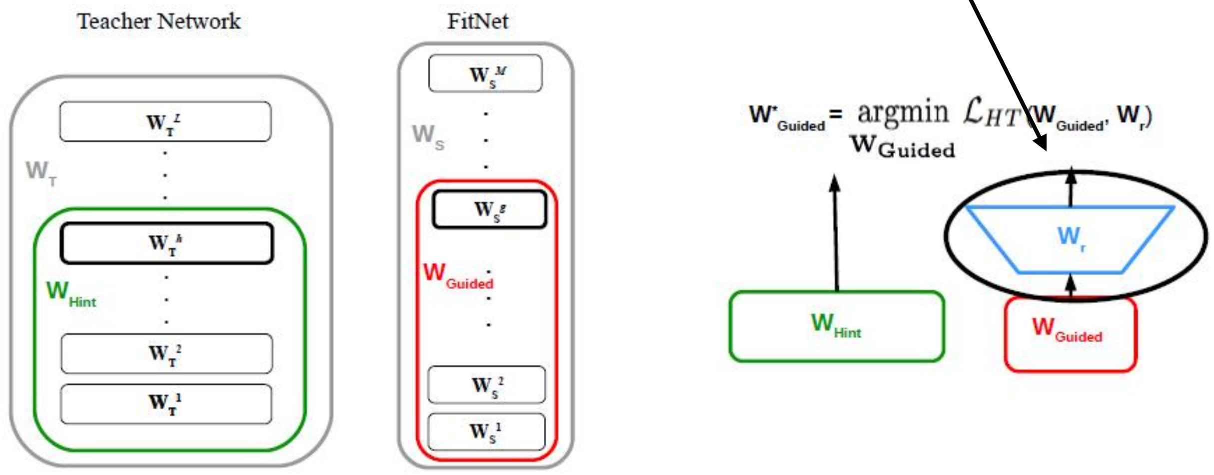
3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation——对齐中间参数

使用FC层对齐参数形状





(a) Teacher and Student Networks (b) Hints Training

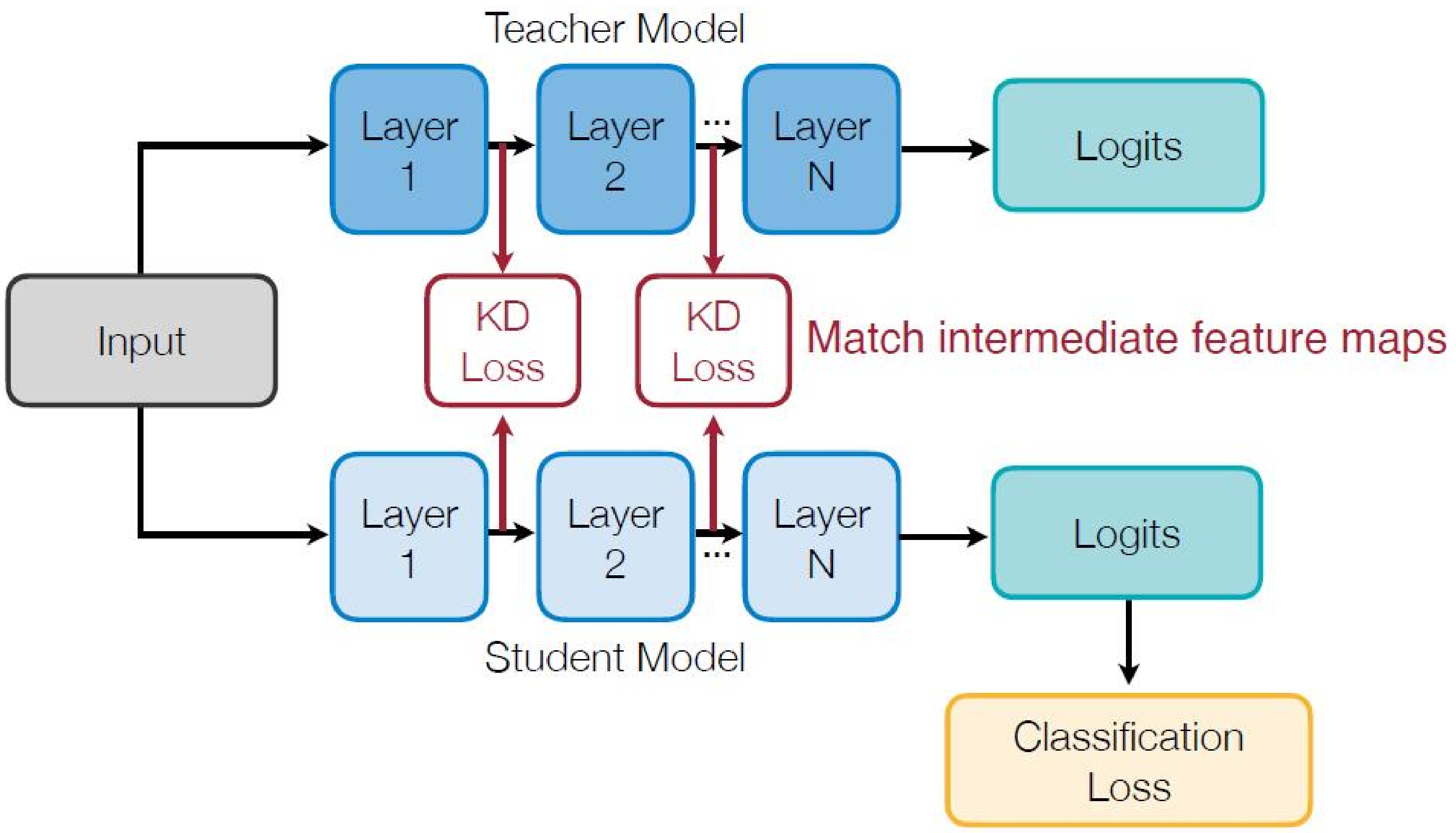
使用L2损失约束教师和学生模型的中间参数

FitNets: Hints for Thin Deep Nets [Romero *et al.*, ICLR 2015]

3.2 基于正则化的类增量学习

Distillation Based

■ Knowledge Distillation——对齐中间特征

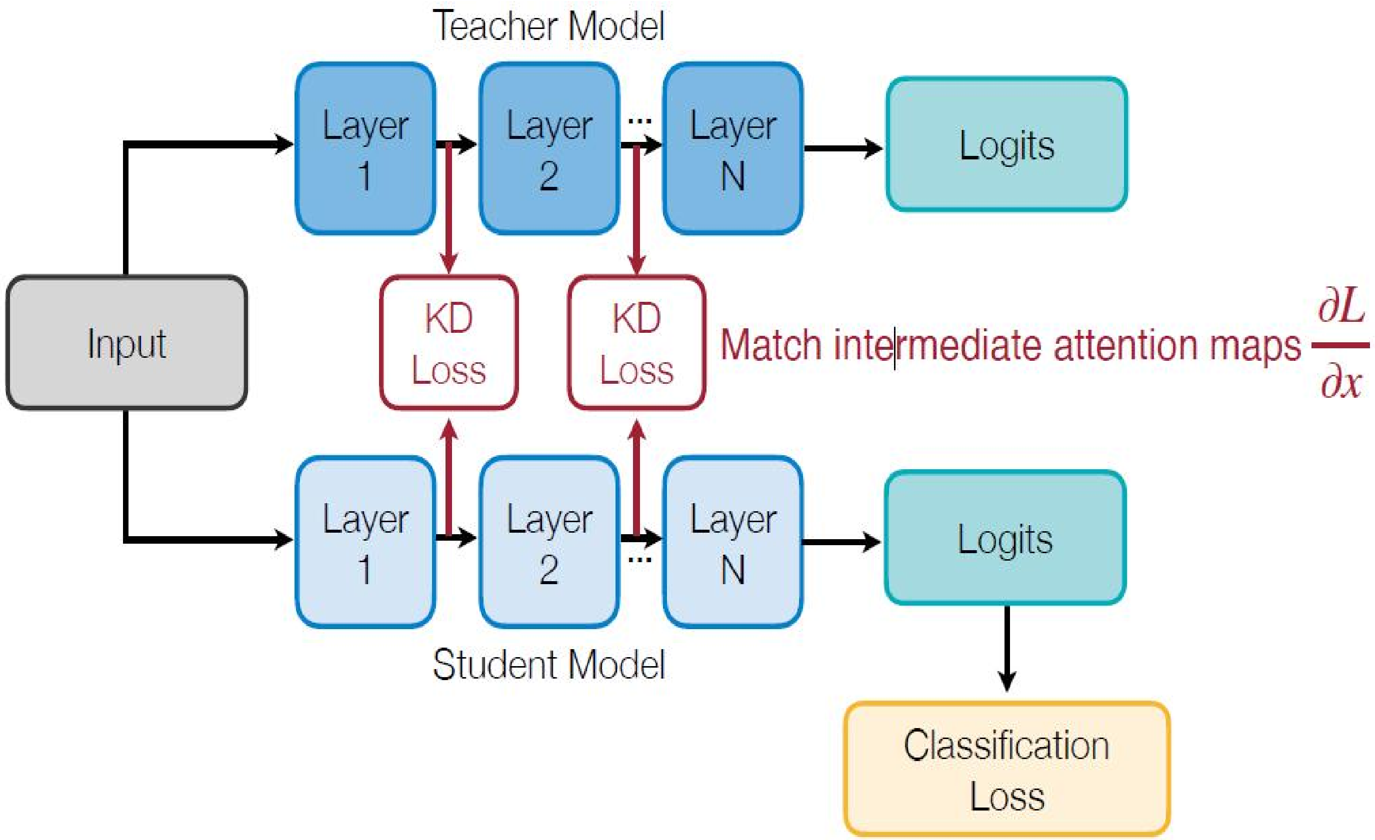


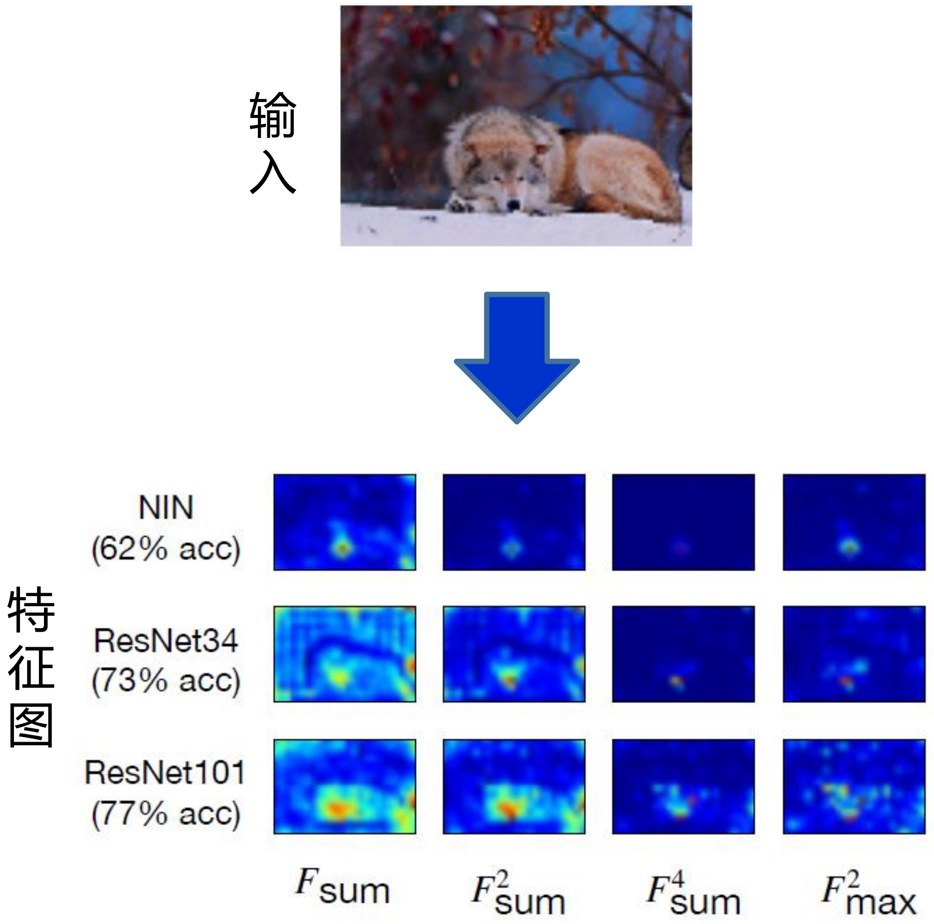
Like What You Like: Knowledge Distill via Neuron Selectivity Transfer [Huang and Wang, arXiv 2017]

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation——对齐注意力图



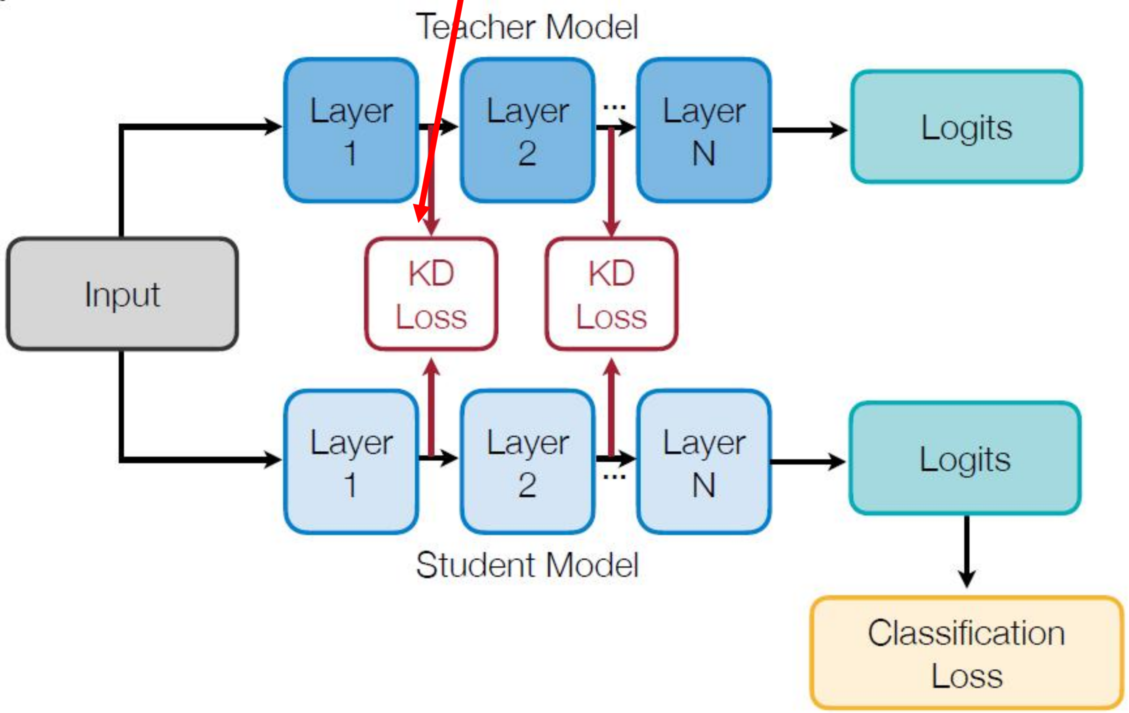


Like What You Like: Knowledge Distill via Neuron Selectivity Transfer [Huang and Wang, arXiv 2017]

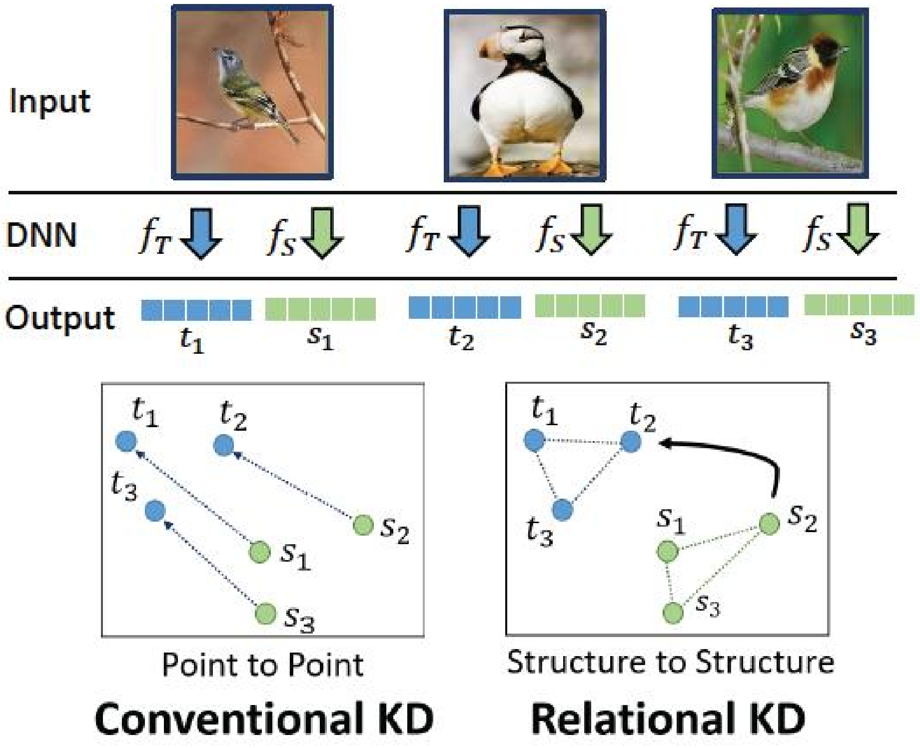
Distillation Based

- Knowledge Distillation——对齐more?

对齐稀疏模式$\rho(x) = 1[x > 0]$



对齐多个样本之间的关系



Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons [Heo et al., AAAI 2019]
Relational Knowledge Distillation [Park *et al.*, CVPR 2019]

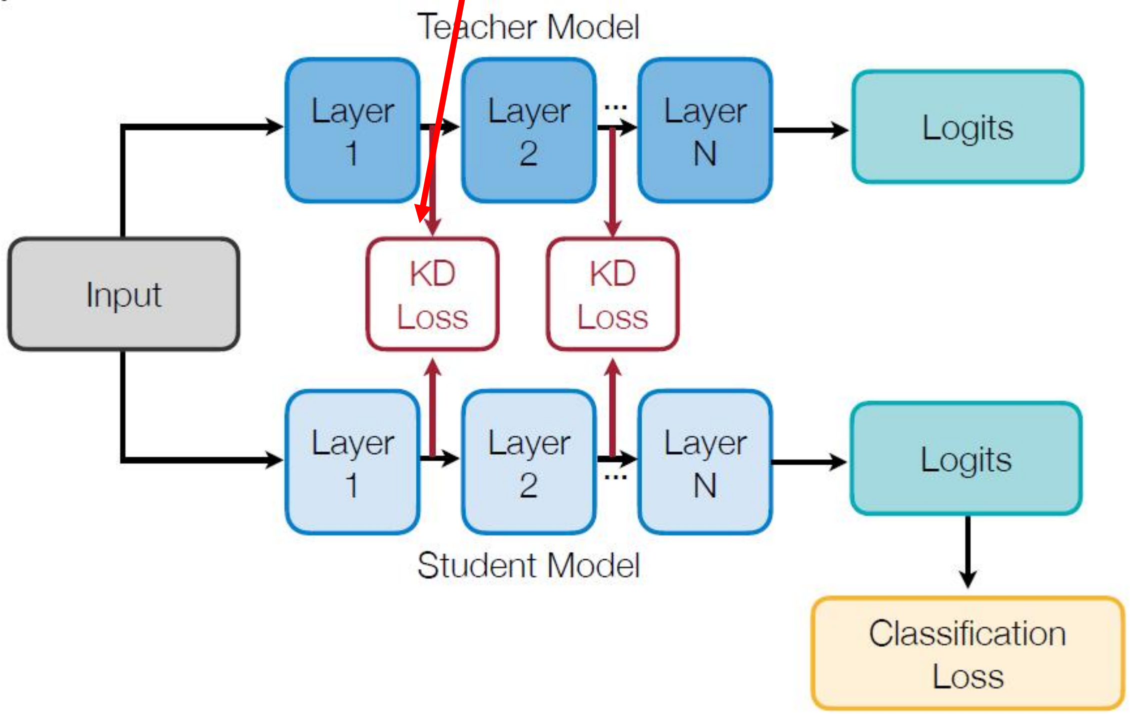
3.2 基于正则化的类增量学习

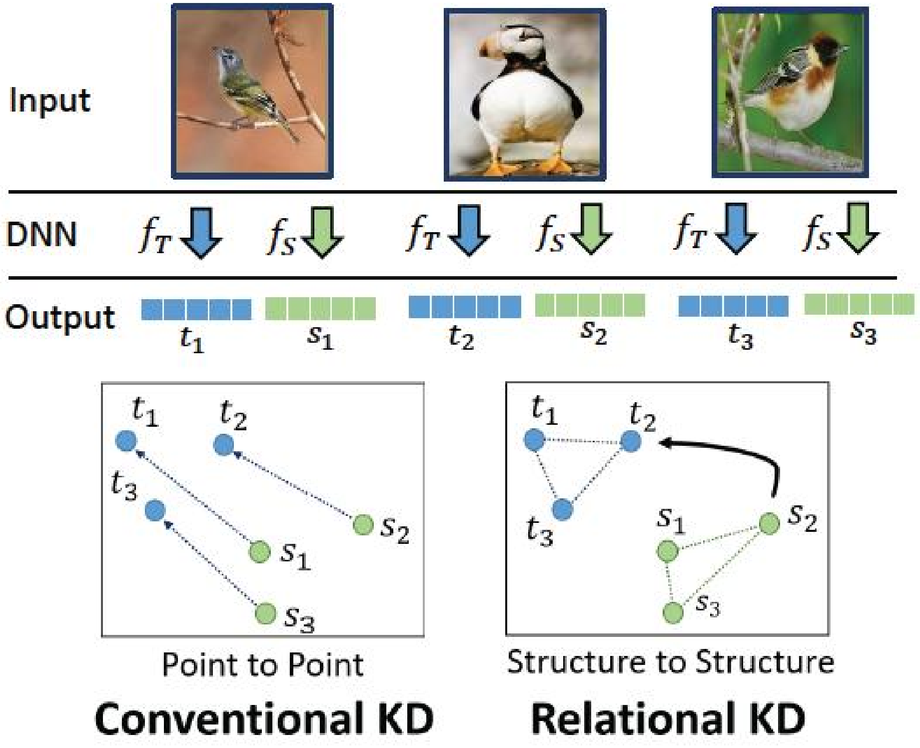
Distillation Based

- Knowledge Distillation——对齐more?

对齐稀疏模式$\rho(x) = 1[x > 0]$

对齐多个样本之间的关系



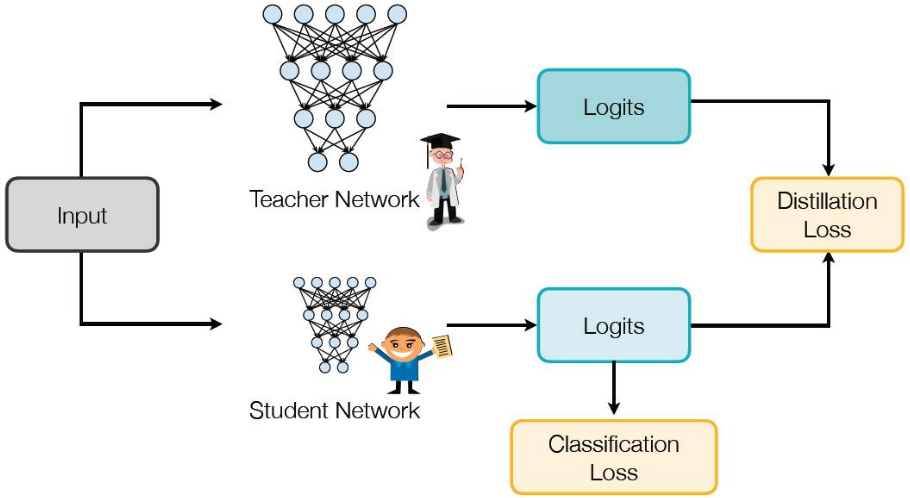


Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons [Heo et al., AAAI 2019]
Relational Knowledge Distillation [Park *et al.*, CVPR 2019]

3.2 基于正则化的类增量学习

Distillation Based

- Knowledge Distillation



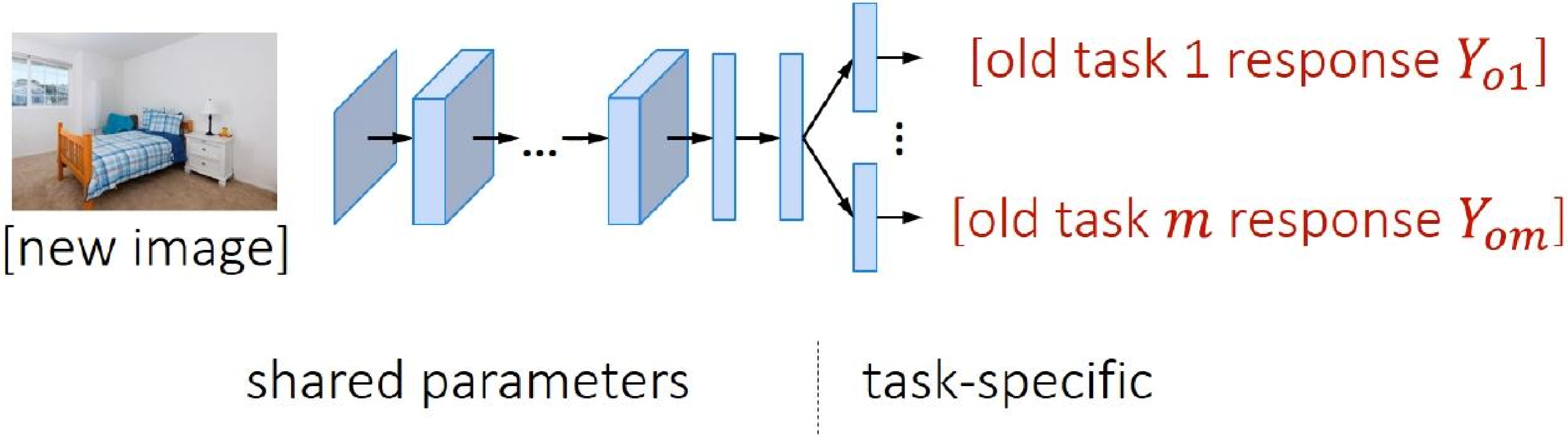
使用KD从旧模型中蒸馏知识用于新任务学习

3.2 基于正则化的类增量学习

Distillation Based

- Learning without Forgetting

  获得旧任务响应



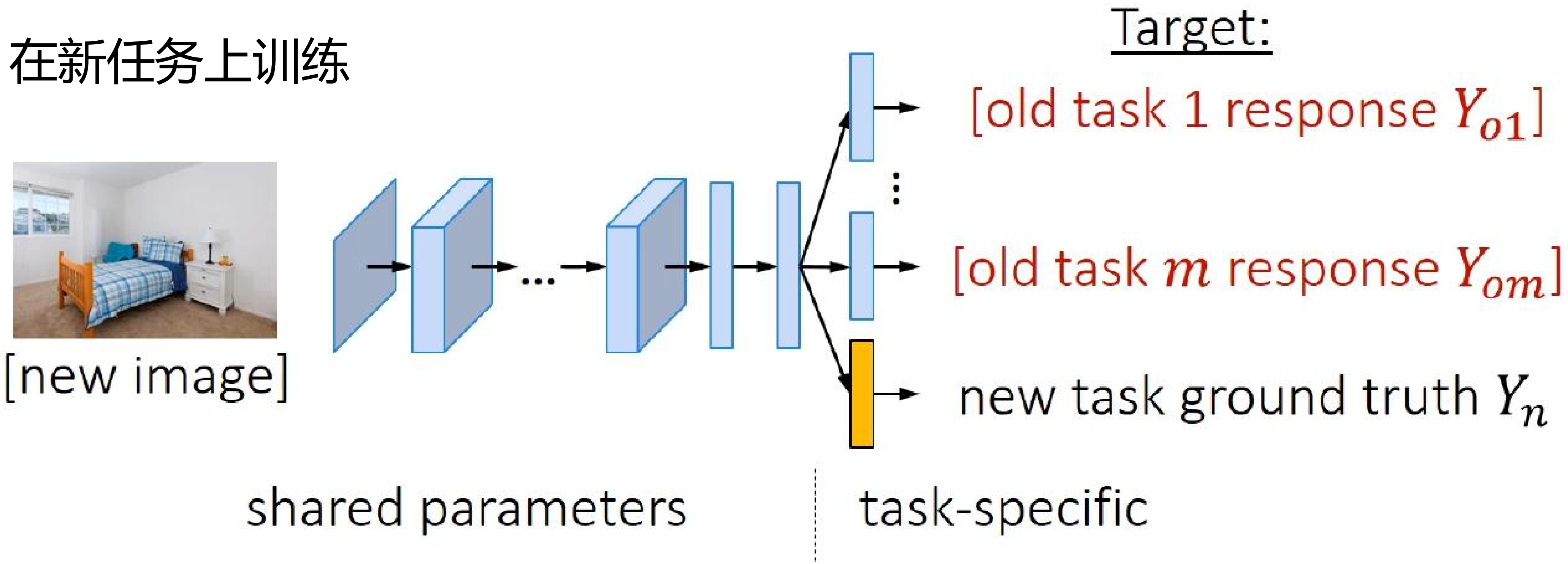
Li, Z., & Hoiem, D. (2017). Learning without forgetting. IEEE PAMI, 40(12), 2935-2947.

3.2 基于正则化的类增量学习

Distillation Based

- Learning without Forgetting

在新任务上训练



$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{old}(Y_{oi}, \hat{Y_{oi}}) + \mathcal{L}_{new}(Y_n, \hat{Y_n}) + \mathcal{R}(\theta)$$

Distillation Based

- Learning without Forgetting

$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{old}(Y_{oi}, \hat{Y_{oi}}) + \mathcal{L}_{new}(Y_n, \hat{Y_n}) + \mathcal{R}(\theta)$$

$$\mathcal{L}_{new}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = -\mathbf{y}_n \cdot \log \hat{\mathbf{y}}_n$$

新任务样本预测与GT的损失

$$\mathcal{L}_{old}(\mathbf{y}_o, \hat{\mathbf{y}}_o) = -H(\mathbf{y}'_o, \hat{\mathbf{y}}'_o) = -\sum_{i=1}^{l} y'^{(i)}_o \log \hat{y}'^{(i)}_o$$

$$y'^{(i)}_o = \frac{(y^{(i)}_o)^{1/T}}{\sum_j (y^{(j)}_o)^{1/T}} \qquad \hat{y}'^{(i)}_o = \frac{(\hat{y}^{(i)}_o)^{1/T}}{\sum_j (\hat{y}^{(j)}_o)^{1/T}}$$

旧模型响应

新模型响应

3.2 基于正则化的类增量学习

Distillation Based

- Learning without Forgetting

LEARNINGWITHOUTFORGETTING:

Start with:

θ_s: shared parameters

θ_o: task specific parameters for each old task

X_n, Y_n: training data and ground truth on the new task

Initialize:

$Y_o \leftarrow \text{CNN}(X_n, \theta_s, \theta_o)$ // *compute output of old tasks for new data*

$\theta_n \leftarrow \text{RANDINIT}(|\theta_n|)$ // *randomly initialize new parameters*

Train:

Define $\hat{Y}_o \equiv \text{CNN}(X_n, \hat{\theta}_s, \hat{\theta}_o)$ // *old task output*

Define $\hat{Y}_n \equiv \text{CNN}(X_n, \hat{\theta}_s, \hat{\theta}_n)$ // *new task output*

$$\theta_s^*,\ \theta_o^*,\ \theta_n^* \leftarrow \underset{\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n}{\operatorname{argmin}} \left(\lambda_o \mathcal{L}_{old}(Y_o, \hat{Y}_o) + \mathcal{L}_{new}(Y_n, \hat{Y}_n) + \mathcal{R}(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n) \right)$$

3.2 基于正则化的类增量学习

Distillation Based

- Learning without Forgetting

AlexNet	1 old task	+	1 new task
	ILSVRC 2012 Places2	+	PASCAL VOC 2012 Caltech-UCSD Birds MIT indoor scenes MNIST

3.2 基于正则化的类增量学习

Distillation Based

- Learning without Forgetting

性能表现（数值为与LWF的差值）

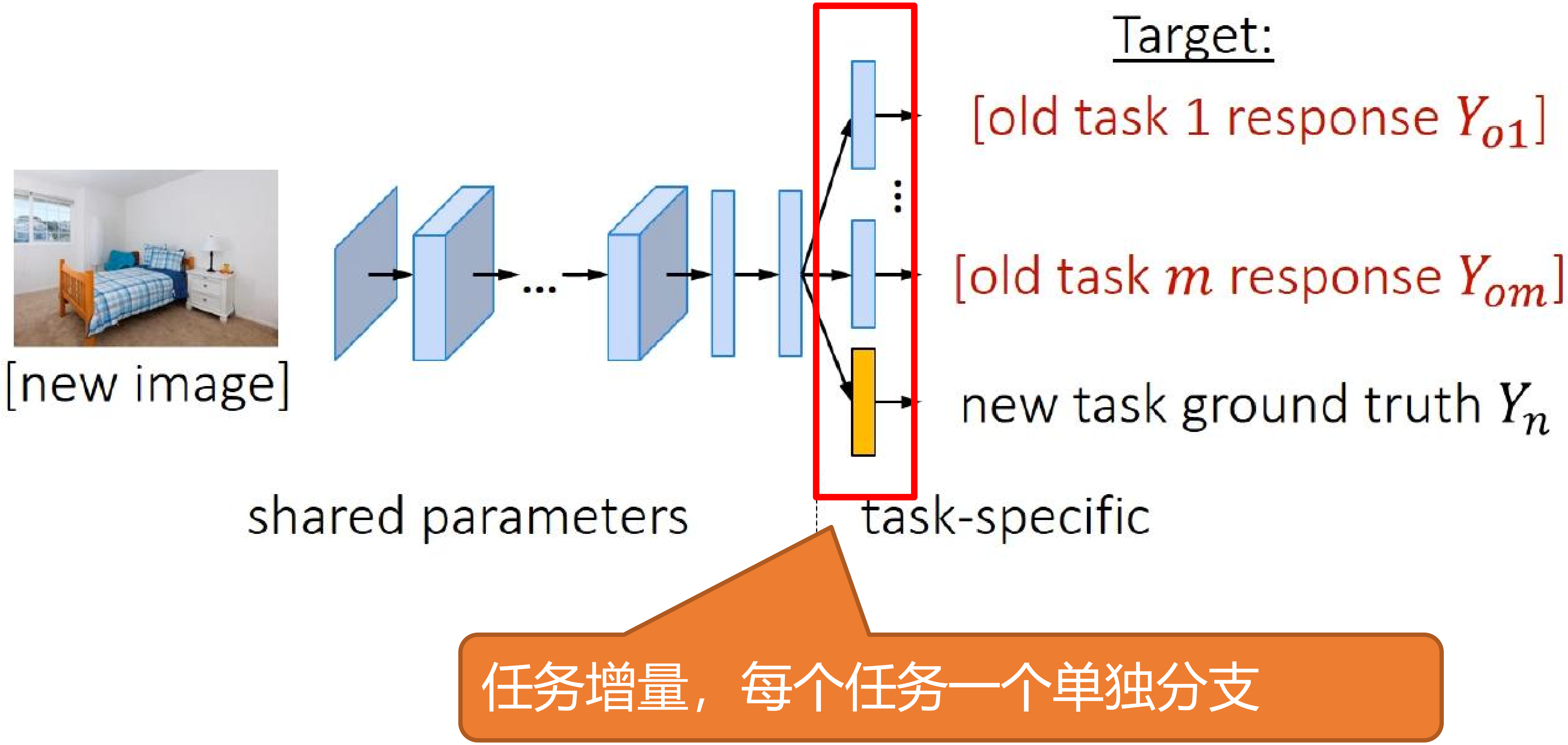
	ImageNet→VOC		ImageNet→CUB		ImageNet→Scenes		ImageNet→MNIST		Places365→VOC		Places365→CUB		Places365→Scenes		Places365→MNIST	
	old	new	old	new	old	new	old	new	old	new	old	new	old	new	old	new
LwF (ours)	56.2	76.1	54.7	57.7	55.9	64.5	49.8	99.3	50.6	70.2	47.9	34.8	50.9	75.2	38.3	99.2
Fine-tuning	-0.9	-0.3	-3.8	-0.7	-2.0	-0.8	-2.8	0.0	-2.2	0.1	-4.6	1.0	-2.1	-1.7	-0.9	0.1
LFL	0.0	-0.4	-1.9	-2.6	-0.3	-0.9	-2.9	-0.6	0.2	-0.7	0.7	-1.7	-0.2	-0.5	-0.4	-0.1
Fine-tune fc	0.5	-0.7	0.2	-3.9	0.6	-2.1	7.0	-0.2	0.5	-1.3	1.8	-4.9	0.3	-1.1	13.0	-0.2
Feat. Extraction	0.8	-0.5	2.3	-5.2	1.2	-3.3	7.3	-0.8	1.1	-1.4	3.8	-12.3	0.8	-1.7	13.3	-1.1
Joint Training	0.7	-0.2	0.6	-1.1	0.5	-0.6	7.2	-0.0	0.7	-0.0	2.3	1.5	0.3	-0.3	13.4	-0.1

新任务上，LWF取得了最好的效果
旧任务上，LWF抄过Fine-tuning，稍差于Joint Training

3.2 基于正则化的类增量学习

Distillation Based

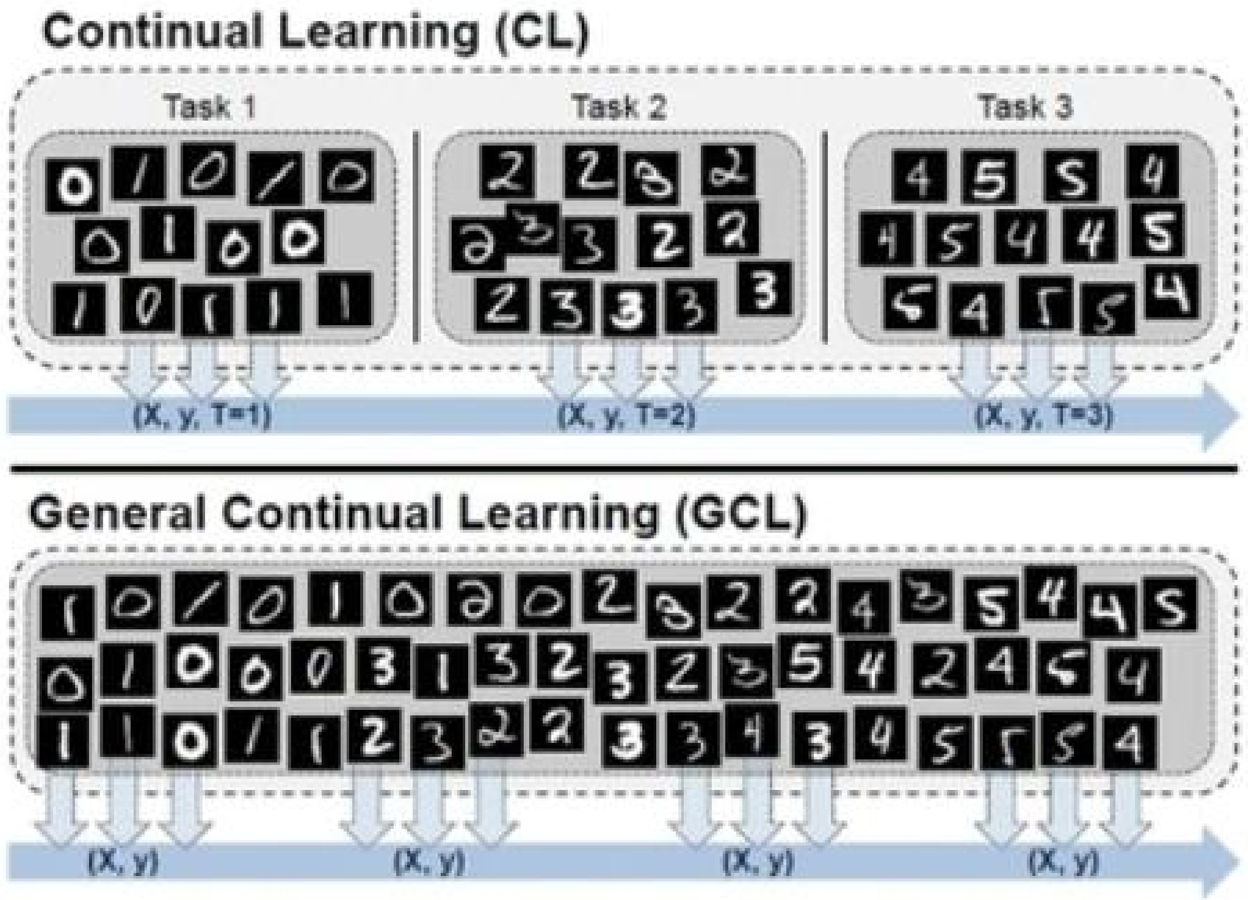
- Learning without Forgetting



3.2 基于正则化的类增量学习

Distillation Based

■ Dark Experience Replay



■ 测试无任务ID

■ 有限存储

Buzzega, Pietro, et al. "Dark experience for general continual learning: a strong, simple baseline." *NIPS 2020*

Distillation Based

- Learning without Forgetting

$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{old}(Y_{oi}, \hat{Y_{oi}}) + \mathcal{L}_{new}(Y_n, \hat{Y_n}) + \mathcal{R}(\theta)$$

新旧数据不平衡

$$\mathcal{L} = \lambda \sum_{i=1}^{m} \mathcal{L}_{old}\,(1-\lambda) + (1-)\mathcal{L}_{new} + \mathcal{R}(\theta)$$

$$\lambda = \frac{|\mathcal{Y}_{b-1}|}{|\mathcal{Y}_b|}$$

旧数据在总样本中的比例

Wu et al, Large scale incremental learning, CVPR 2019

3.2 基于正则化的类增量学习

Distillation Based

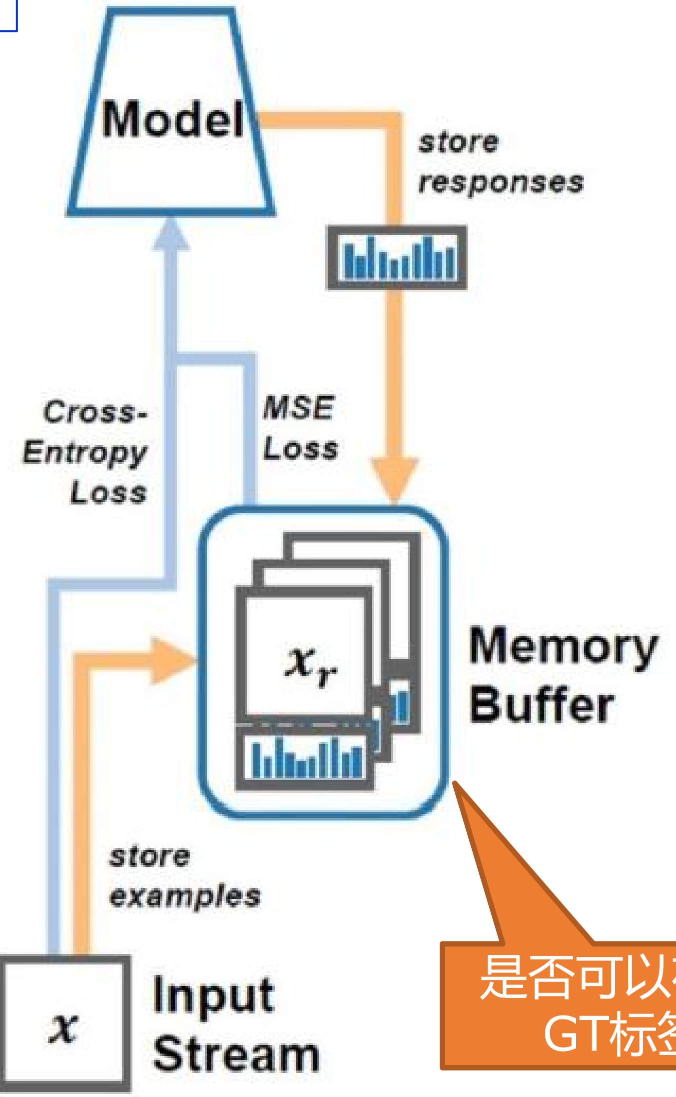
- Dark Experience Replay
  - 使用Buffer存储部分样本响应(Logits)
  - 最小化现在输出与Buffer存储logit的L2距离

当前任务CE损失

$$\underset{\theta}{\text{argmin}}\, \mathcal{L}_{t_c} + \alpha\, \mathbb{E}_{(x,z)\sim\mathcal{M}}\left[\, \|z - h_\theta(x)\|_2^2 \,\right]$$

$$\mathcal{L}_t \triangleq \mathbb{E}_{(x,y)\sim D_t}\left[\ell(y, f_\theta(x))\right]$$

当前任务CE损失



是否可以存储GT标签

3.2 基于正则化的类增量学习

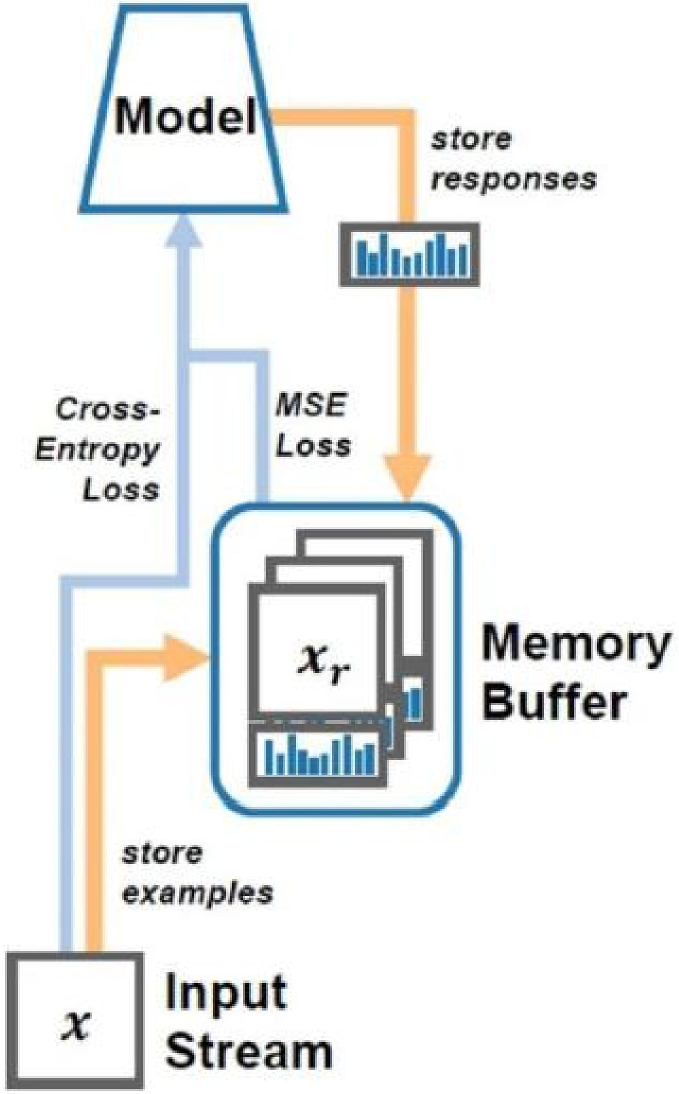
Distillation Based

- Dark Experience Replay
  - 使用Buffer存储部分样本响应(Logits)
  - 最小化现在输出与Buffer存储logit的L2距离

$$\mathcal{L}_{t_c} + \alpha\, \mathbb{E}_{(x',y',z')\sim\mathcal{M}}\left[\,\|z' - h_\theta(x')\|_2^2\,\right]$$

$$+\ \beta\, \mathbb{E}_{(x'',y'',z'')\sim\mathcal{M}}\left[\ell(y'', f_\theta(x''))\right]$$

Buffer样本的GT损失



3.2 基于正则化的类增量学习

Distillation Based

- Dark Experience Replay

Algorithm 1 - Dark Experience Replay

Input: dataset D, parameters θ, scalar α,
learning rate λ

$\mathcal{M} \leftarrow \{\}$
for (x, y) **in** D **do**
  $(x', z', y') \leftarrow sample(\mathcal{M})$
  $x_t \leftarrow augment(x)$
  $x'_t \leftarrow augment(x')$
  $z \leftarrow h_\theta(x_t)$
  $reg \leftarrow \alpha \left\| z' - h_\theta(x'_t) \right\|_2^2$
  $\theta \leftarrow \theta + \lambda \cdot \nabla_\theta [\ell(y, f_\theta(x_t)) + reg]$
  $\mathcal{M} \leftarrow reservoir(\mathcal{M}, (x, z))$
end for

Distillation Based

- Dark Experience Replay

Algorithm 2 - Dark Experience Replay ++

Input: dataset D, parameters θ, scalars α and β, learning rate λ

$\mathcal{M} \leftarrow \{\}$

for (x, y) **in** D **do**

$(x', z', y') \leftarrow sample(\mathcal{M})$

$(x'', z'', y'') \leftarrow sample(\mathcal{M})$

$x_t \leftarrow augment(x)$

$x'_t,\ x''_t \leftarrow augment(x'),\ augment(x'')$

$z \leftarrow h_\theta(x_t)$

$reg \leftarrow \alpha \, \|z' - h_\theta(x'_t)\|_2^2 + \beta \, \ell(y'', f_\theta(x''_t))$

$\theta \leftarrow \theta + \lambda \cdot \nabla_\theta[\ell(y, f_\theta(x_t)) + reg]$

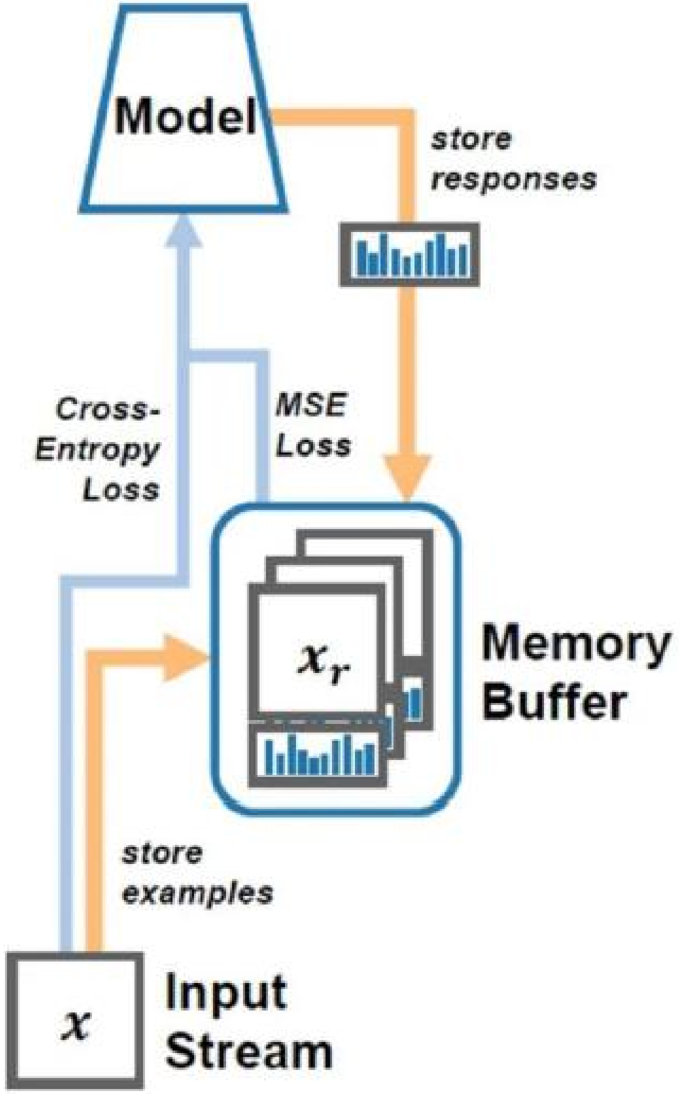
$\mathcal{M} \leftarrow reservoir(\mathcal{M}, (x, z, y))$

end for

3.2 基于正则化的类增量学习

Distillation Based

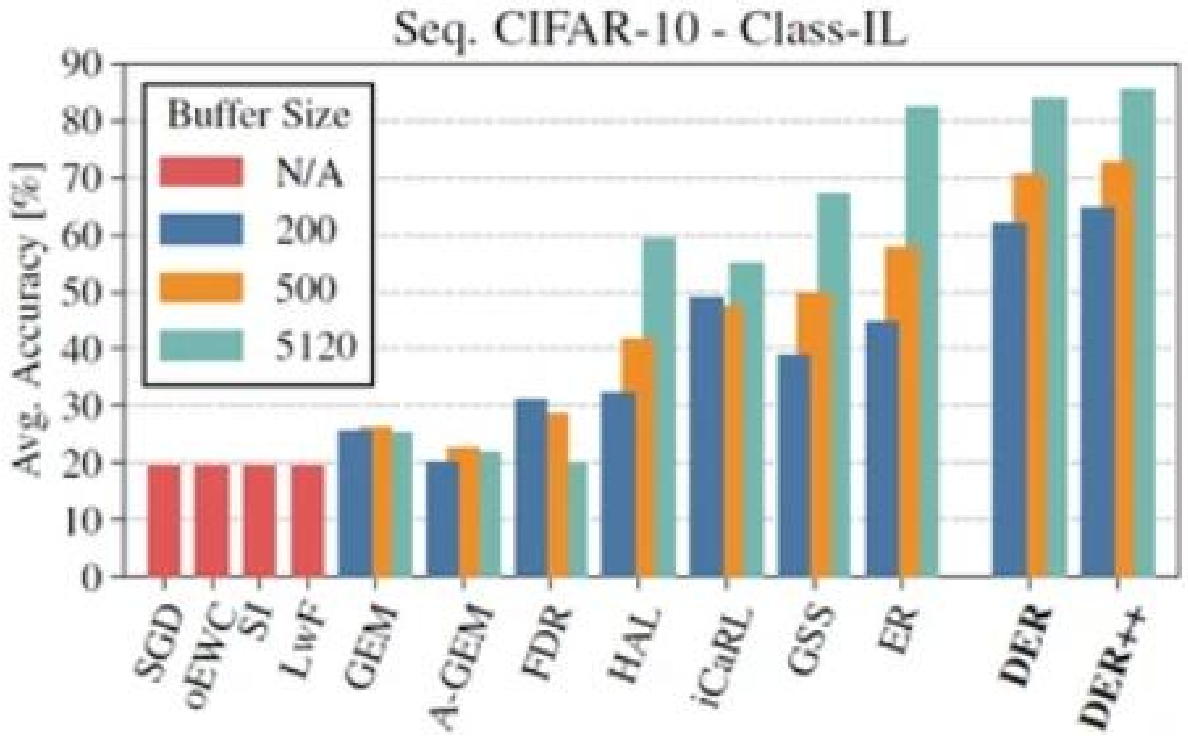
- Dark Experience Replay
  - 使用Buffer存储部分样本响应(Logits)
  - 最小化现在输出与Buffer存储logit的L2距离
  - 在整个优化过程中都更新Buffer

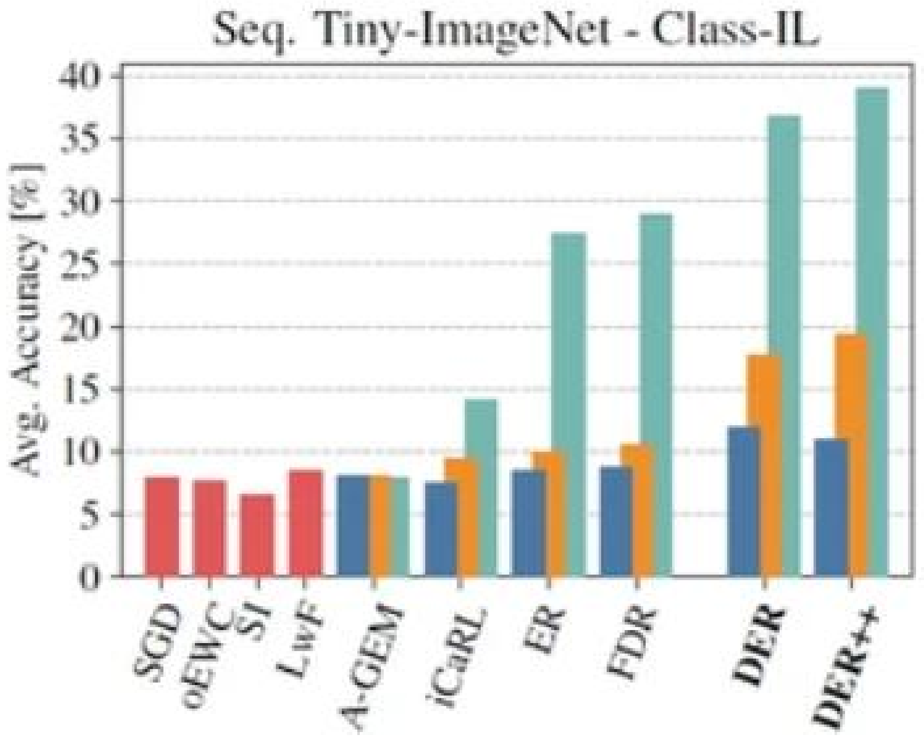


3.2 基于正则化的类增量学习

Distillation Based

- Dark Experience Replay





3.2 基于正则化的类增量学习

Distillation Based

- Dark Experience Replay
  - 优点

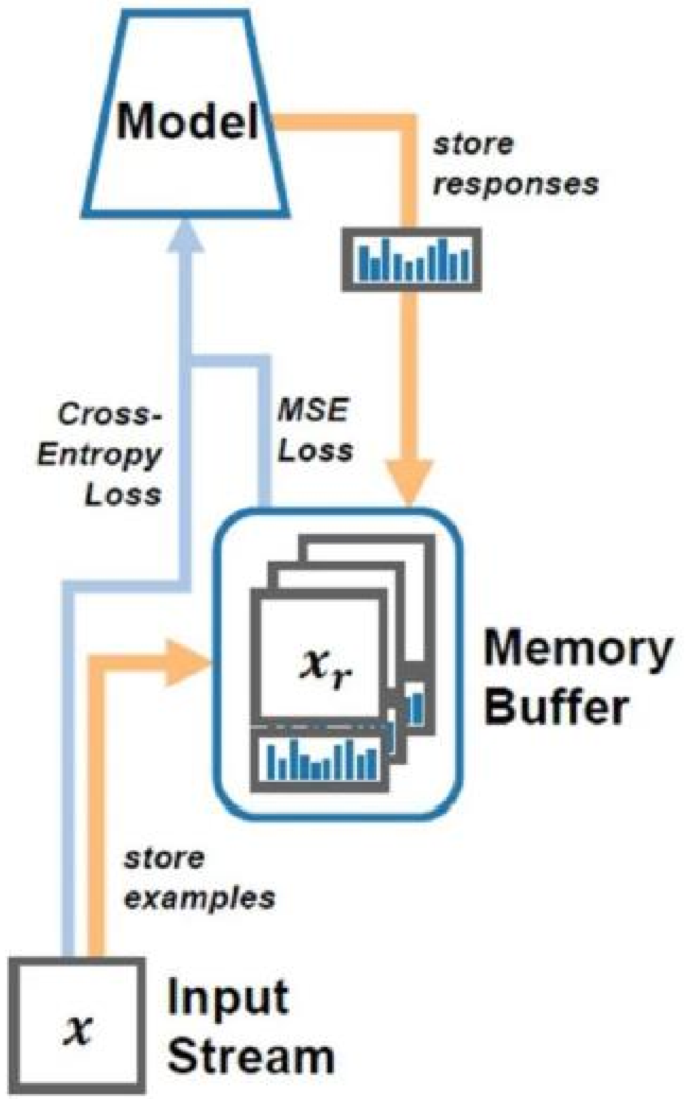
    高效

    测试时候，不需要任务ID
  - 缺点

    蒸馏存在知识损失

    如何平衡新旧知识

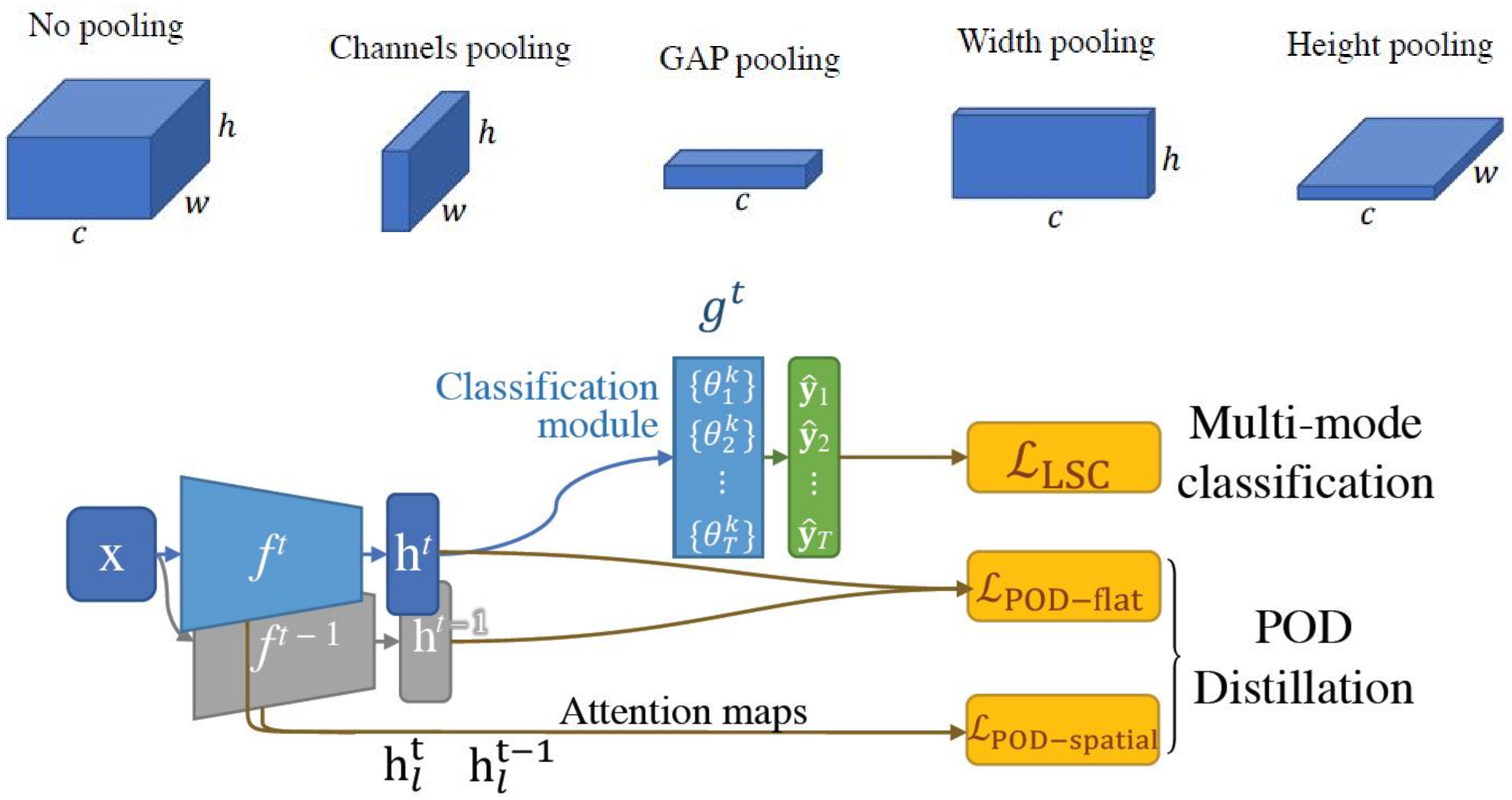
    难以控制实际参数的变化



3.2 基于正则化的类增量学习

Distillation Based

- 特征蒸馏PODNet



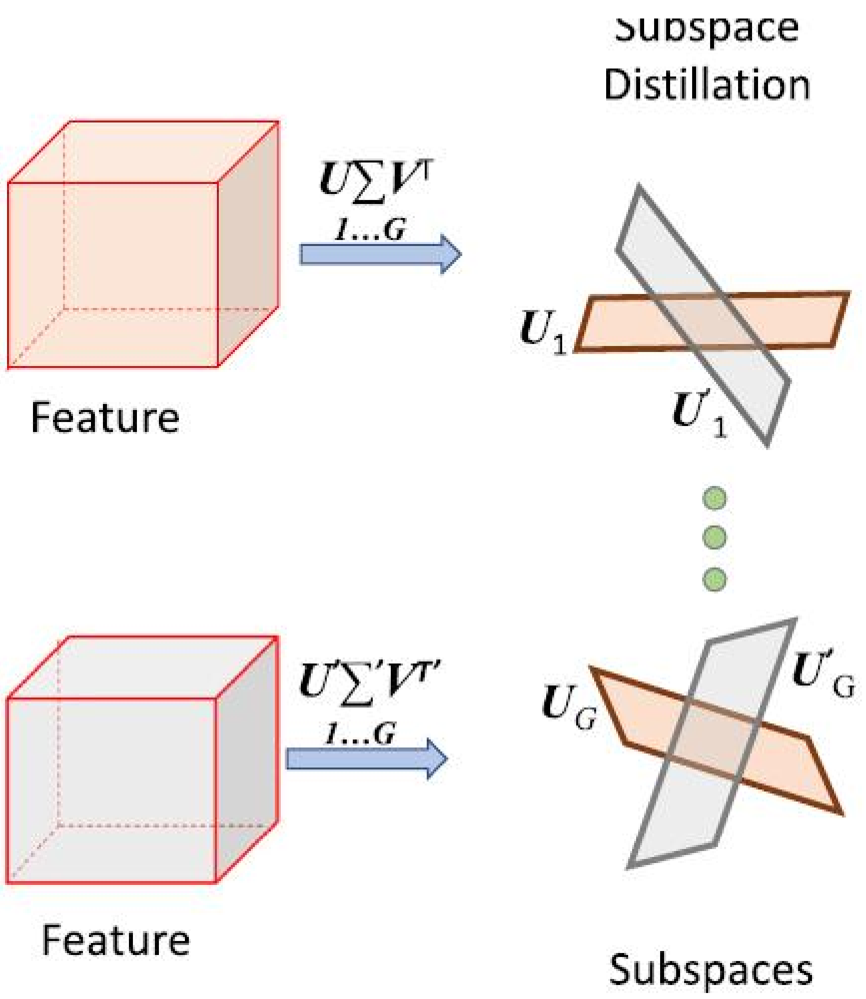
Douillard, Arthur, et al. "Podnet: Pooled outputs distillation for small-tasks incremental learning." *ECCV 2020*

Distillation Based

- 子空间蒸馏

$$\ell_{\mathrm{SD}}^{\mathrm{CL}}(\mathcal{X}_B, \mathcal{Y}_B) := \frac{1}{|\mathbb{C}^t|} \sum_{k=1}^{|\mathbb{C}^t|} \left(2\, m - 2 \left\| \mathbf{P}_k^{t\top} \mathbf{P}_k^{t-1} \right\|_F^2 \right)$$

新旧提取特征的SVD分解



Roy, Kaushik, et al. "Subspace distillation for continual learning." Neural Networks 167 (2023): 65-79.

3.2 基于正则化的类增量学习

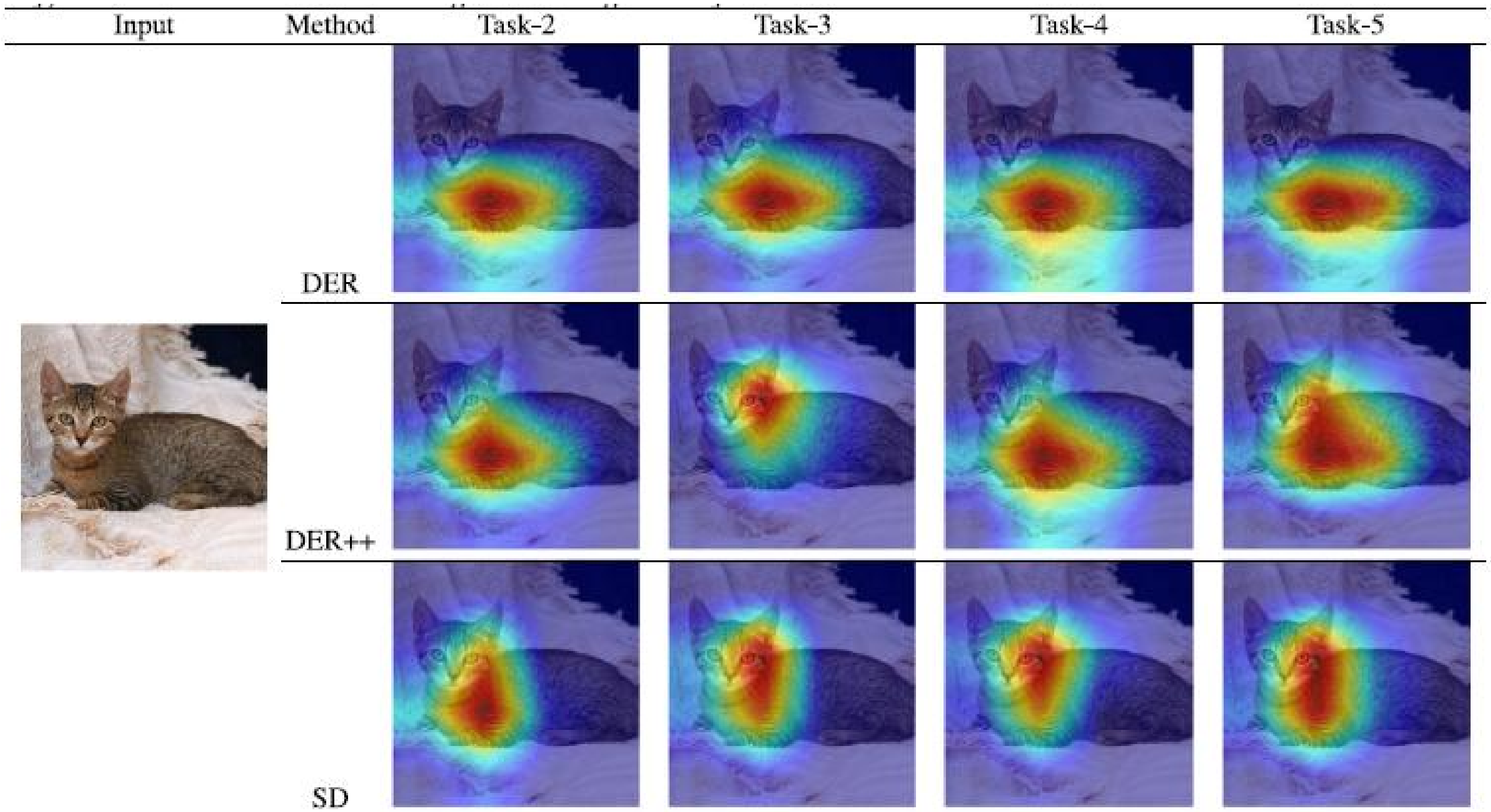
Distillation Based

- Subspace

Method	S-MNIST		S-CIFAR10		S-Tiny Imagenet	
	Task-IL	Class-IL	Task-IL	Class-IL	Task-IL	Class-IL
JOINT	99.65	97.92	98.29	92.20	82.04	59.87
SGD	87.15	19.90	61.02	19.61	17.93	7.79
LwF (Li & Hoiem, 2017)	99.25	20.07	63.28	19.59	15.79	8.46
oEWC (Schwarz et al., 2018)	99.10	20.00	68.27	19.47	19.20	7.56
SI (Zenke et al., 2017)	99.07	19.97	68.05	19.46	35.97	6.58
	Tiny Memory					
ER (Rolnick et al., 2019)	97.72	73.80	77.85	32.87	28.07	5.85
DER (Buzzega et al., 2020)	98.48	77.12	80.72	32.43	27.73	4.26
SD (Ours)	98.35	**79.37**	**81.65**	**35.1**	**30.11**	**6.05**
	Small Memory					
iCARL (Rebuffi et al., 2017)	98.28	70.51	88.99	49.02	28.19	7.53
ER (Rolnick et al., 2019)	97.86	80.43	91.19	44.79	38.17	8.49
DER (Buzzega et al., 2020)	98.80	84.55	91.40	61.93	40.22	11.87
SD (Ours)	97.71	85.28	**92.88**	61.85	39.52	8.54
DER (Buzzega et al., 2020) + SD (Ours)	**98.86**	**86.54**	92.07	**66.12**	**42.63**	**12.26**
	Medium Memory					
iCARL (Rebuffi et al., 2017)	98.81	74.55	88.22	47.55	31.55	9.38
ER (Rolnick et al., 2019)	98.89	86.57	93.61	57.74	48.64	9.99
DER (Buzzega et al., 2020)	98.84	90.54	93.40	70.51	51.78	17.75
SD (Ours)	**99.00**	89.00	**94.86**	71.85	48.60	10.03
DER (Buzzega et al., 2020) + SD (Ours)	98.98	**91.47**	94.68	**75.96**	**52.74**	**19.43**

Distillation Based

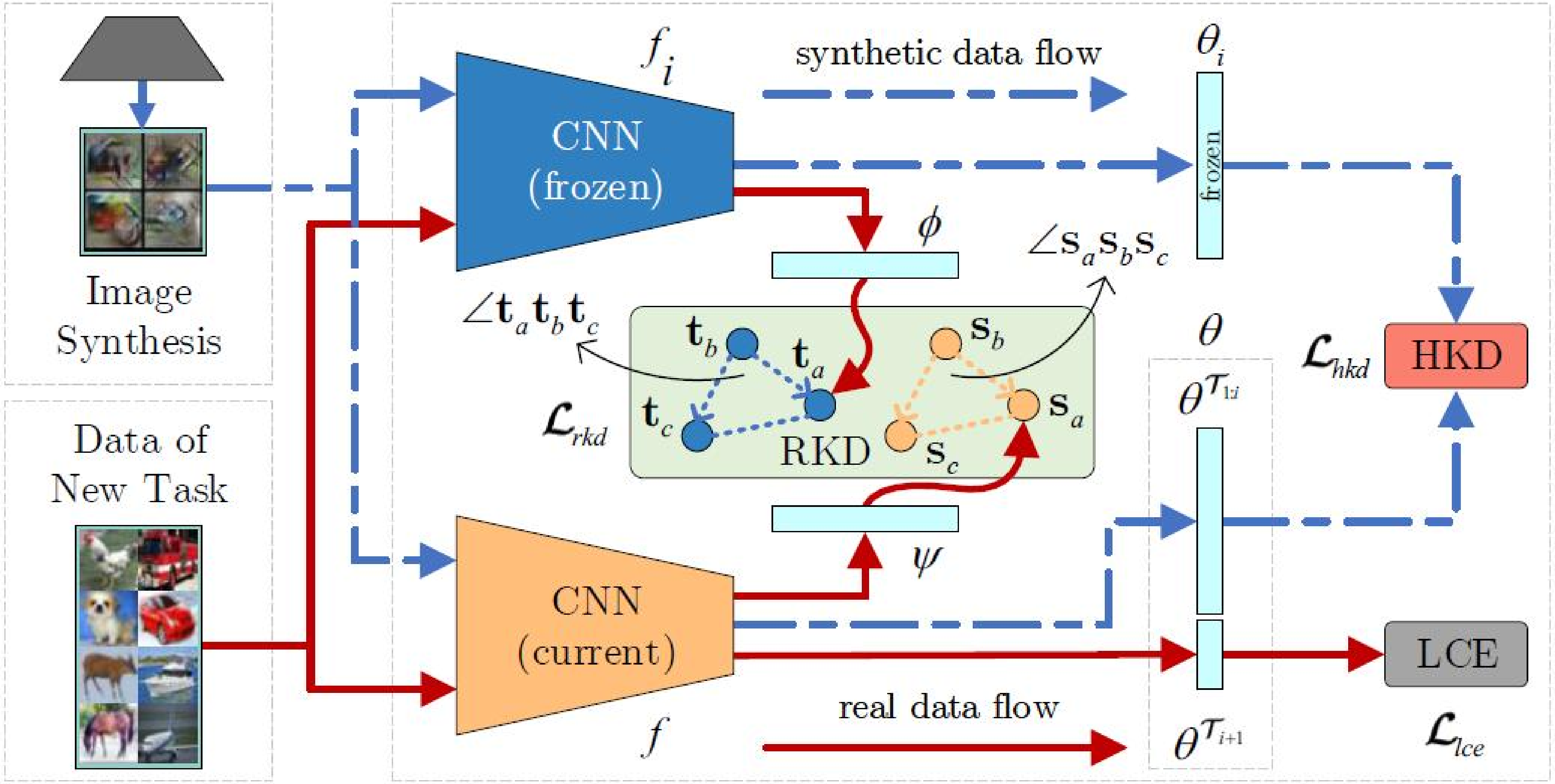
- Subspace distillation



3.2 基于正则化的类增量学习

Distillation Based

- 关系蒸馏

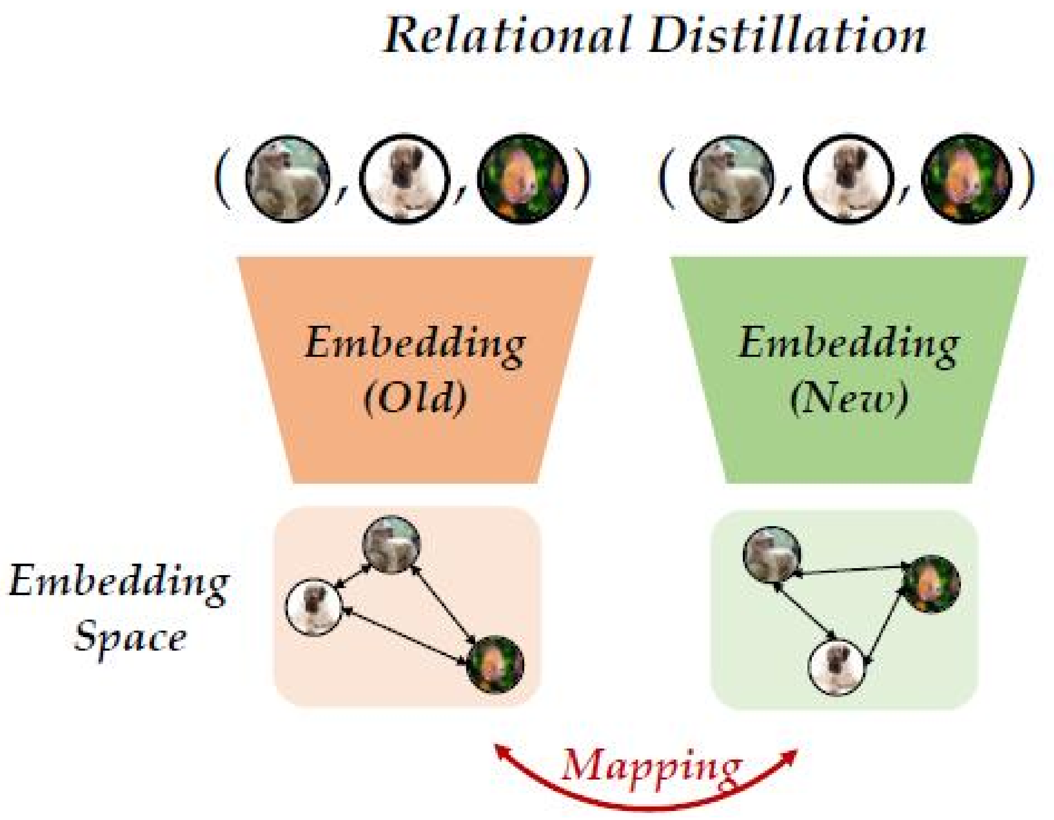


Gao, Qiankun, et al. "R-dfcil: Relation-guided representation learning for data-free class incremental learning." ECCV 2022.

3.2 基于正则化的类增量学习

Distillation Based

- 关系蒸馏



构建三元组 $\{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k\}$

$$\cos\angle\mathbf{r}_a\mathbf{r}_b\mathbf{r}_c = \langle \mathbf{e}^{ab}, \mathbf{e}^{cb} \rangle \qquad \mathbf{e}^{ij} = \frac{\mathbf{r}_i - \mathbf{r}_j}{\|\mathbf{r}_i - \mathbf{r}_j\|_2}$$

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

旧模型表征

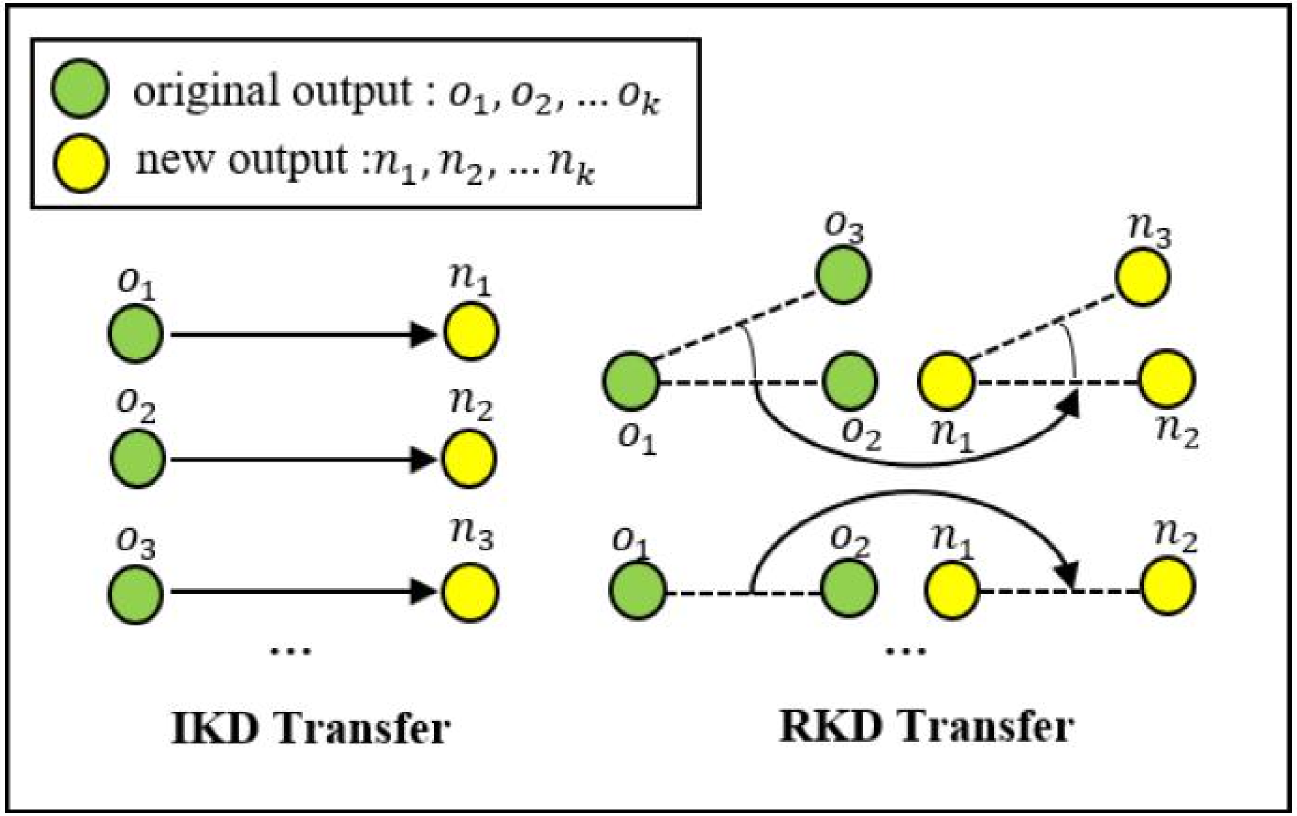
新模型的表征

Gao, Qiankun, et al. "R-dfcil: Relation-guided representation learning for data-free class incremental learning." ECCV 2022.

3.2 基于正则化的类增量学习

Distillation Based

- 关系蒸馏



构建Exemplar Relation Graph

$$A(p,q,z;\Theta^t) = \langle e_{pq}, e_{zq} \rangle,\ p,q,z \subset G^t$$

$$e_{pq} = \frac{v_p - v_q}{\|v_p - v_q\|_2},\quad e_{zq} = \frac{v_z - v_q}{\|v_z - v_q\|_2}$$

$$A(p,q,z;\Theta^{t+1}) = \langle e_{pq}, e_{zq} \rangle,\ p,q,z \subset G^t$$

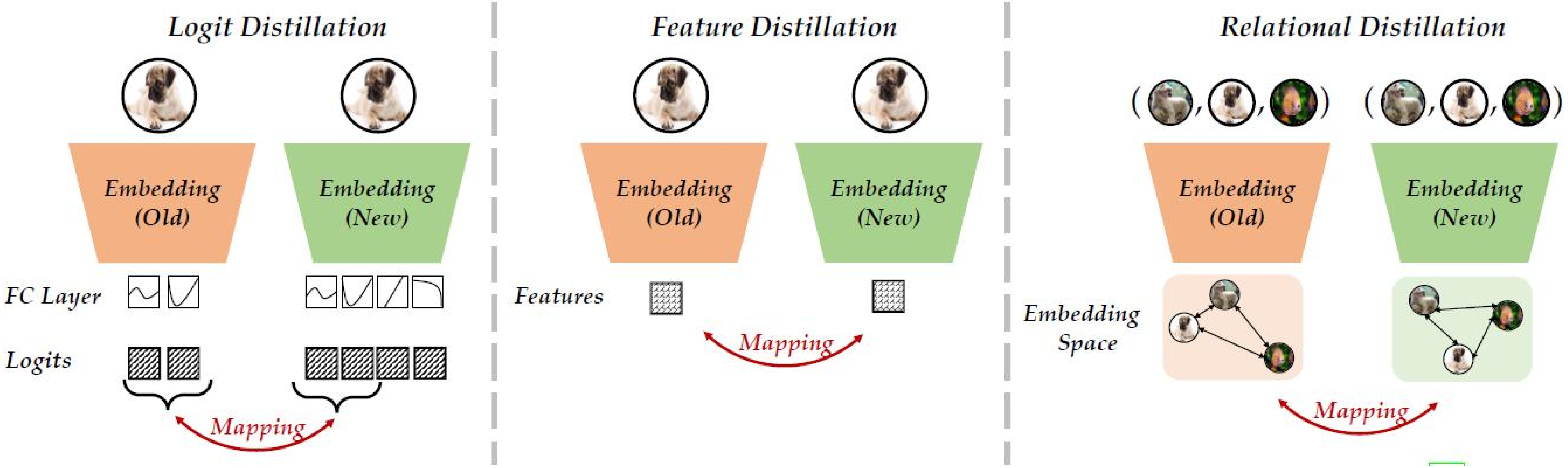
where $e_{pq} = \frac{v_p - v_q}{\|v_p - v_q\|_2},\quad e_{zq} = \frac{v_z - v_q}{\|v_z - v_q\|_2}$

$$\ell_{ERL}(G^t;\Theta^t,\Theta^{t+1}) = |A(\Theta^t) - A(\Theta^{t+1})|_p$$

3.2 基于正则化的类增量学习

Distillation Based

■ 总结

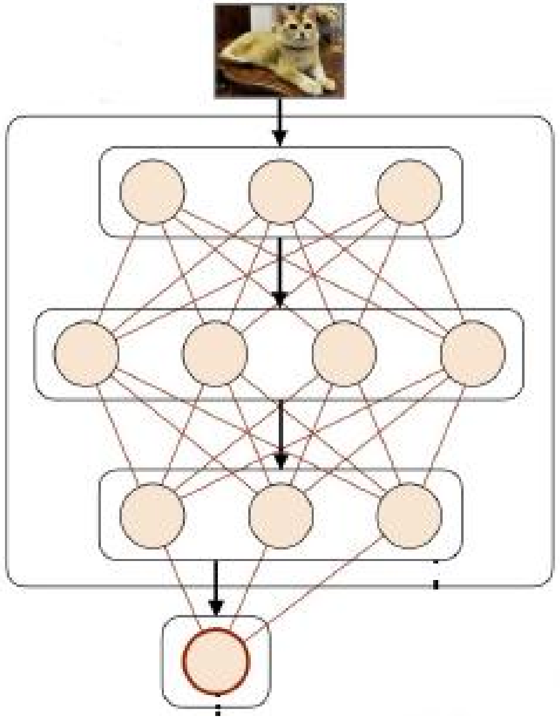


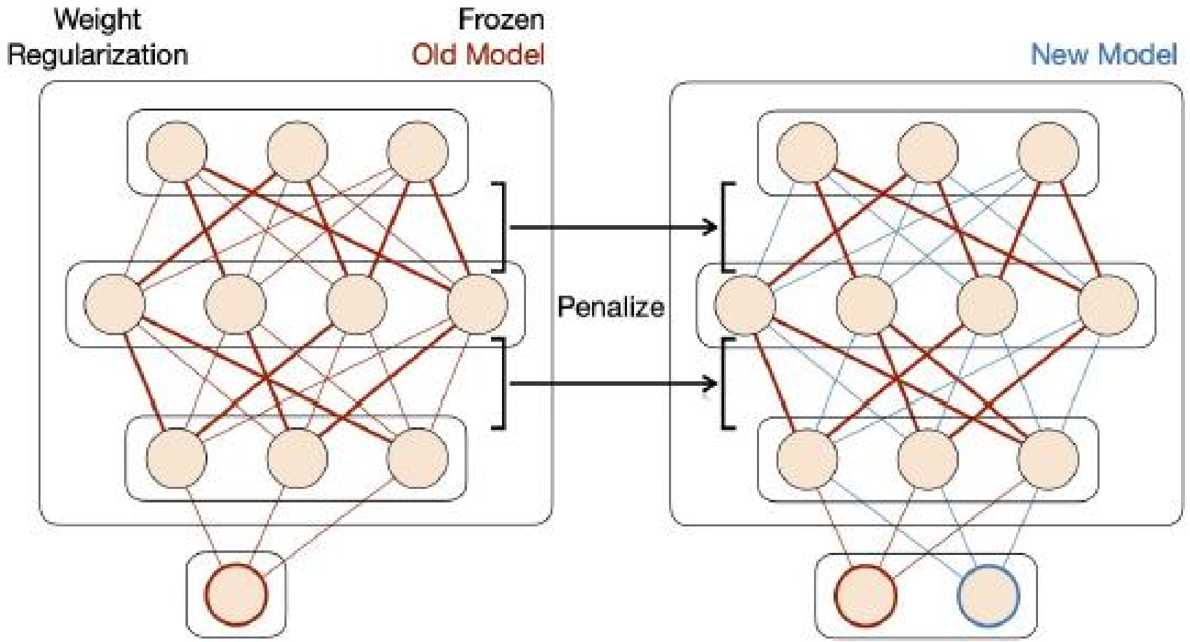
Zhou et al, Deep Class-Incremental Learning: A Survey

3.2 基于正则化的类增量学习

基于参数重要性

- 考虑参数对任务贡献是不同的
- 估计重要性分布，作为先验优化指导模型学习

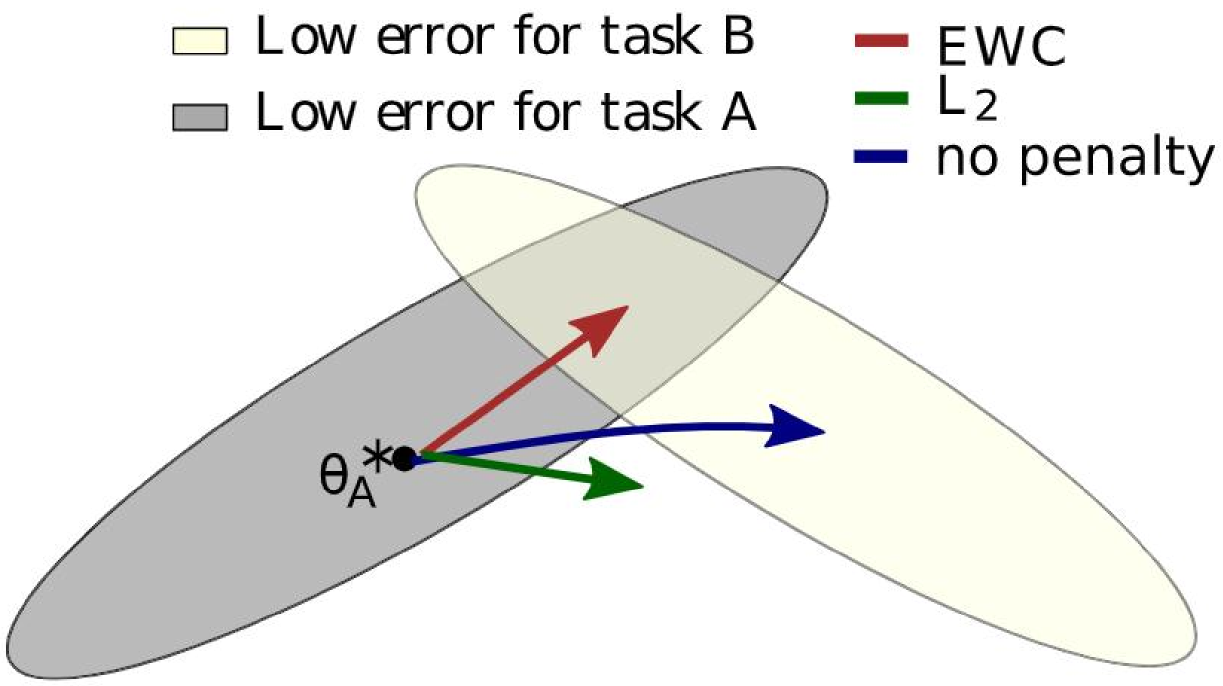




3.2 基于正则化的类增量学习

基于参数重要性

- EWC
  - ✓ 标志性的参数重要性方法



Don't let important parameters change drastically (reduce plasticity)

James Kirkpatrick, et al. Overcoming catastrophic forgetting in neural networks. PNAS 2017

3.2 基于正则化的类增量学习

基于参数重要性

- EWC
  - ✓ 标志性的参数重要性方法

通用性公式，不同是如何计算参数重要性矩阵

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,i})^2$$

当前任务的BCE损失

参数重要性矩阵

旧模型参数

3.2 基于正则化的类增量学习

基于参数重要性

- EWC
  - ✓ 标志性的参数重要性方法

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,i})^2$$

Fisher信息矩阵
损失的梯度幅值，越大代表越重要

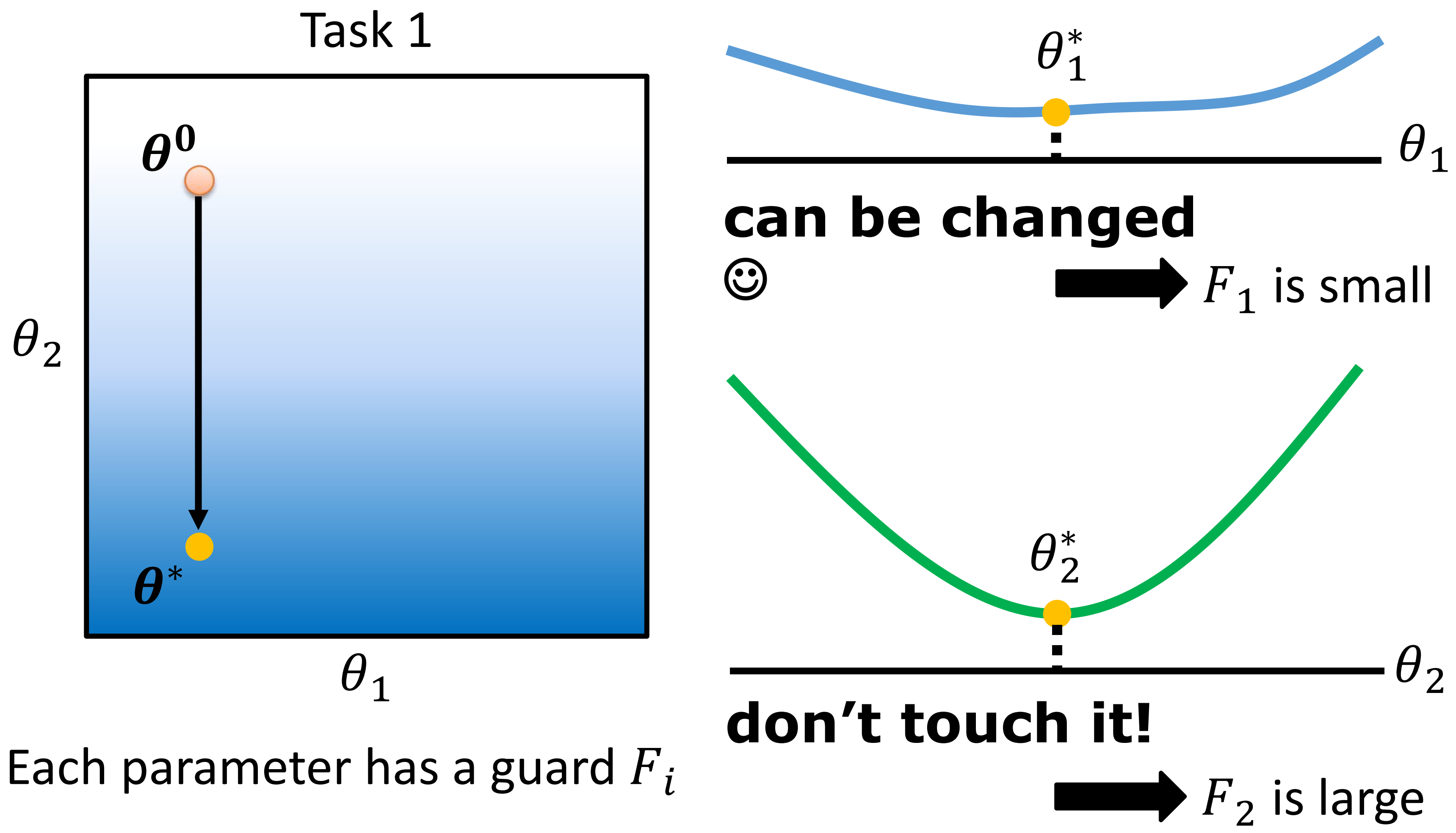
$$F_\theta = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\left(\frac{\partial \log p_\theta(\mathbf{y}|\mathbf{x})}{\partial \theta}\right)\left(\frac{\partial \log p_\theta(\mathbf{y}|\mathbf{x})}{\partial \theta}\right)^\top\right]$$

- ✓ Fisher 信息矩阵等于对数似然函数的海森矩阵的期望取负
- ✓ 反映了对数似然函数在参数处的曲率
- ✓ 曲率越大，对数似然函数越高而窄，否则越平而宽

3.2 基于正则化的类增量学习

基于参数重要性

- EWC

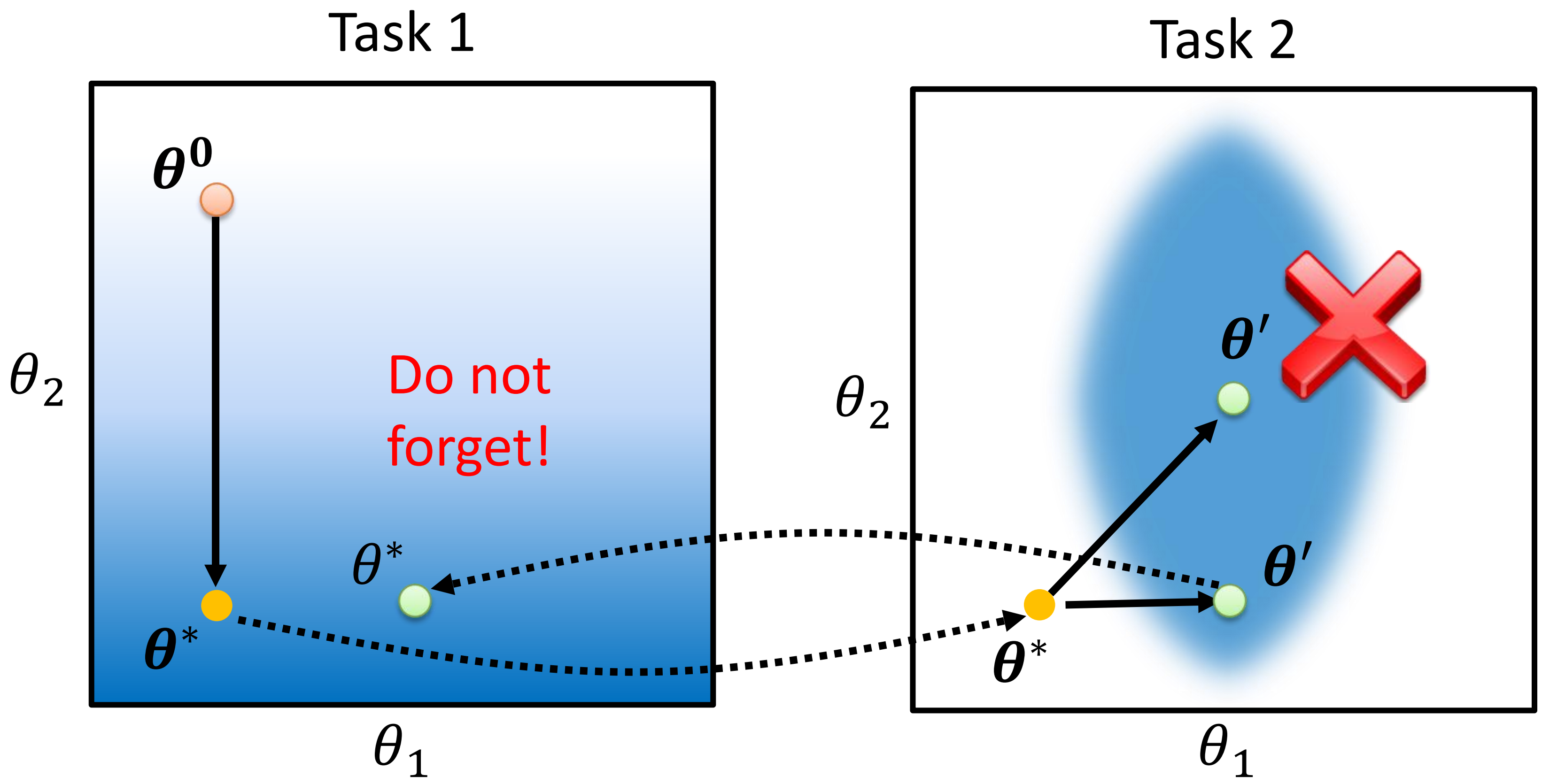


Cited From Li hongyi's ML Course

3.2 基于正则化的类增量学习

基于参数重要性

- EWC



F_1 is small, while F_2 is large.

(We can modify θ_1, but do not change θ_2.)

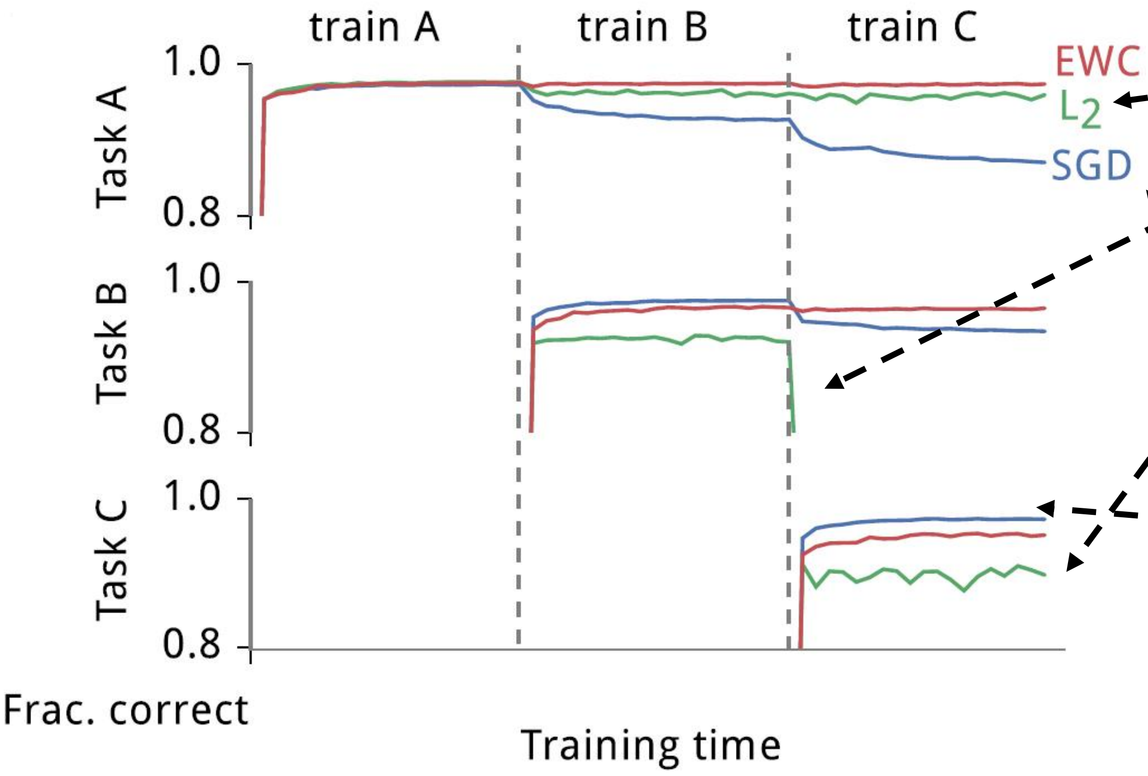
Cited From Li hongyi's ML Course

3.2 基于正则化的类增量学习

基于参数重要性

- EWC

  ✓性能



L2 is too rigid, doesn't allow learning on new tasks

$F_i = 1$

Catastrophic Forgetting in SGD

$F_i = 0$

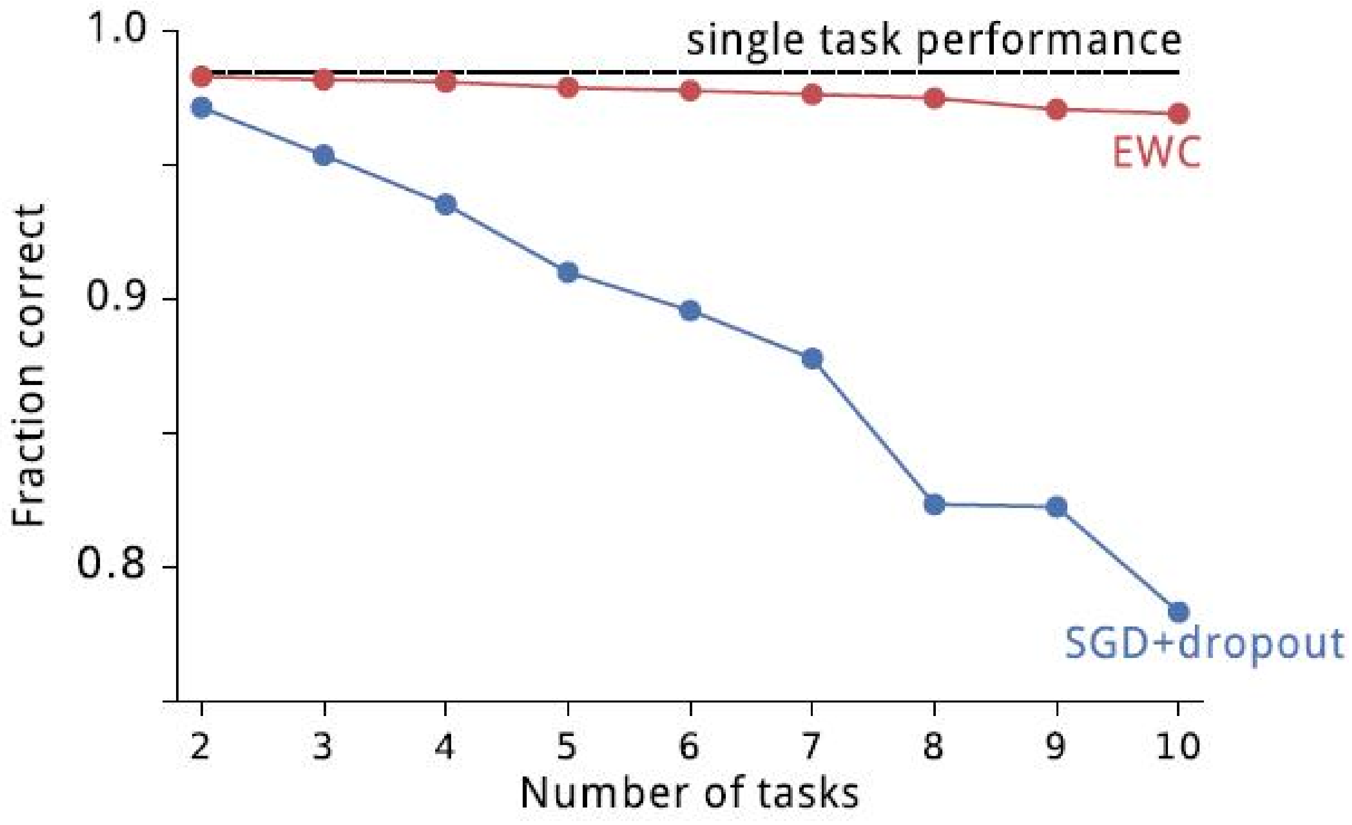
MNIST permutation, from the original EWC paper

3.2 基于正则化的类增量学习

基于参数重要性

- EWC

  ✓性能



3.2 基于正则化的类增量学习

基于参数重要性

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

- EWC++：按批次迭代更新

$$F_\theta^t = \alpha F_\theta^t + (1-\alpha) F_\theta^{t-1}$$

在当前批数据上计算

3.2 基于正则化的类增量学习

基于参数重要性

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

- EWC++：按批次迭代更新

$$F_\theta^t = \alpha F_\theta^t + (1-\alpha) F_\theta^{t-1}$$

参数 θ_k 对任务 ν 的重要性：
1.沿着训练轨迹上对任务v整体损失的贡献 ω_k^ν；
2.该参数的移动距离 Δ_k^ν

- Synaptic Intelligence：按批次迭代更新

单个参数k对任务v的贡献

$$\tilde{L}_\mu = L_\mu + c\sum_k \Omega_k^\mu \left(\tilde{\theta}_k - \theta_k\right)^2 \qquad \Omega_k^\mu = \sum_{\nu<\mu} \frac{\omega_k^\nu}{(\Delta_k^\nu)^2 + \xi}$$

$$\sum_t g_k(\theta(t))\theta'_k(t) \qquad \boldsymbol{g} = \frac{\partial L}{\partial \boldsymbol{\theta}}$$

$$\Delta_k^\nu \equiv \theta_k(t^\nu) - \theta_k(t^{\nu-1})$$

Continual Learning Through Synaptic Intelligence, ICML 2018

3.2 基于正则化的类增量学习

基于参数重要性

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,i})^2$$

✓Elastic Weight Consolidation (EWC)
https://arxiv.org/abs/1612.00796

✓Synaptic Intelligence (SI)
https://arxiv.org/abs/1703.04200

✓Memory Aware Synapses (MAS)
https://arxiv.org/abs/1711.09601

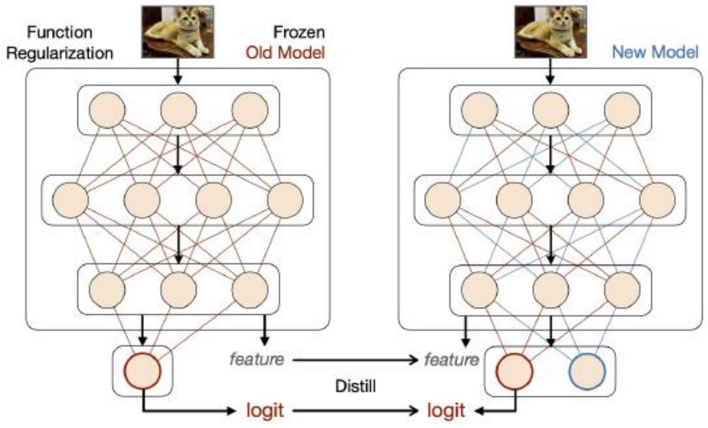
✓RWalk
https://arxiv.org/abs/1801.10112

✓Sliced Cramer Preservation (SCP)
https://openreview.net/forum?id=BJge3TNKwH

3.2 基于正则化的类增量学习

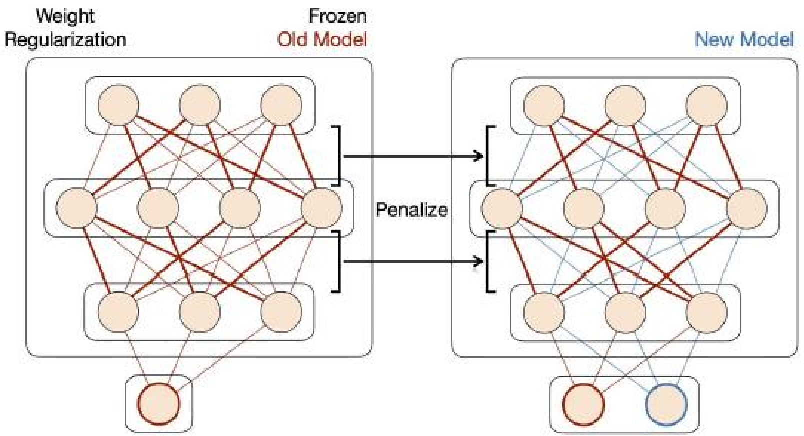
总结　实际优化 $\frac{1}{N_{\mathcal{T}}}\sum_{i=1}^{N_{\mathcal{T}}}\ell(f(x_i^{(\mathcal{T})};\theta),y_i^{(\mathcal{T})})$ +penalty Term

Distillation



可塑性效果不明显

先验模型（参数重要性）



度量矩阵占用存储空间

如何快速计算度量是个问题

谢　　谢！