

Reusable

模型的复用与可复用性探究

Reuse

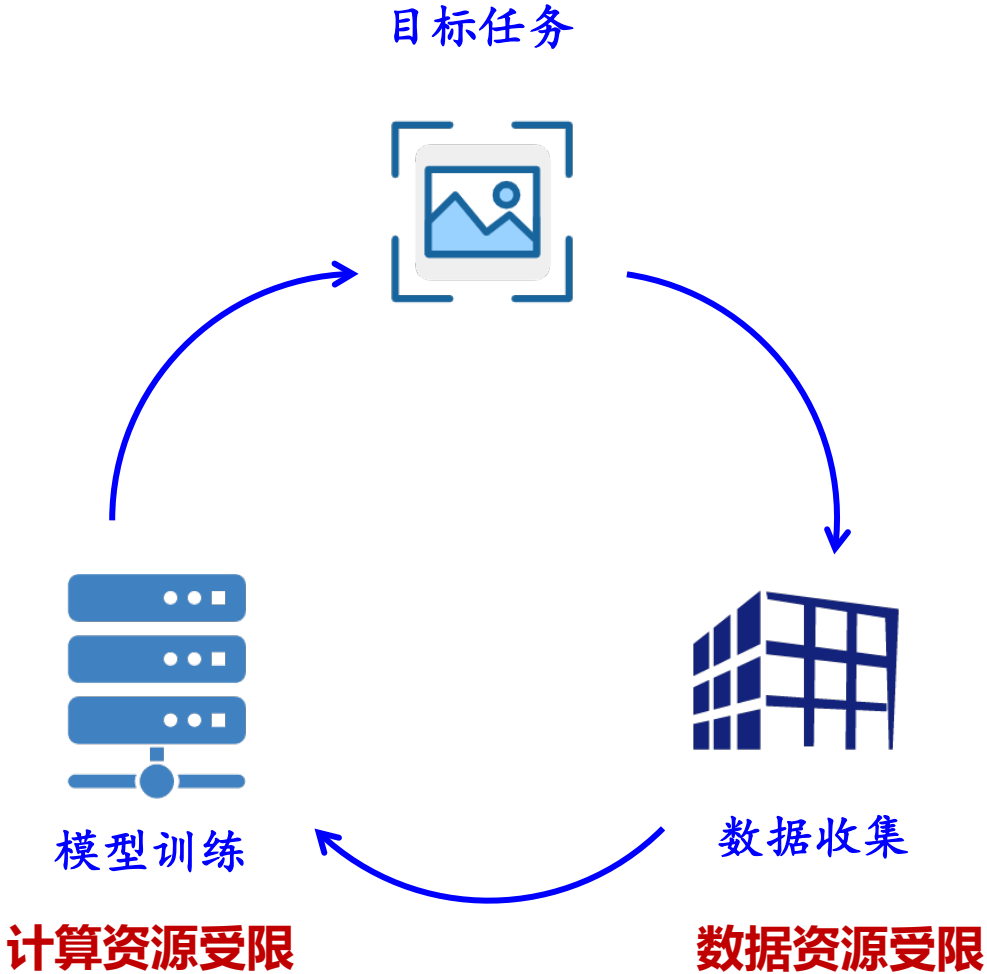
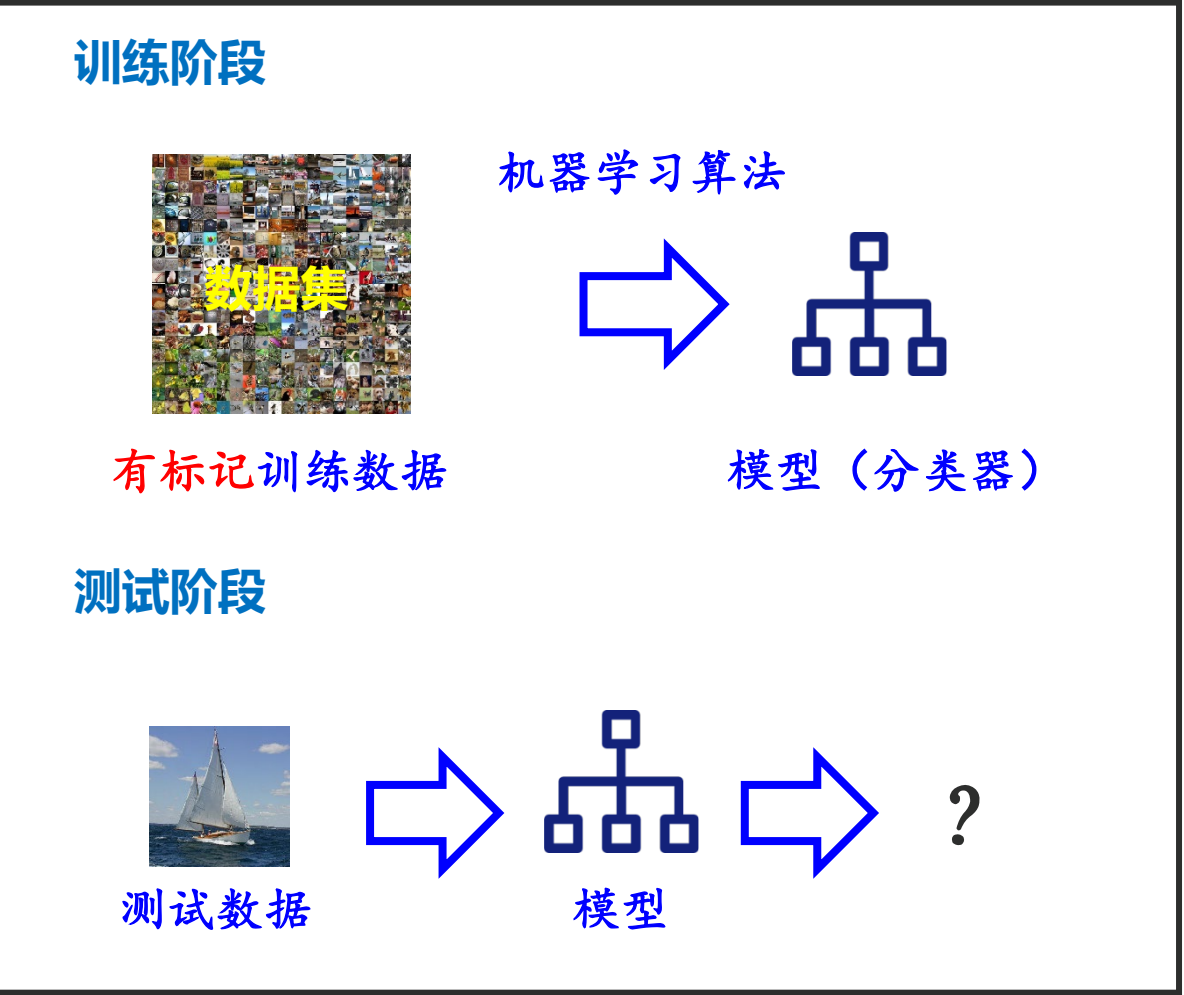
叶翰嘉

南京大学

yehj@nju.edu.cn



监督学习中的数据和模型



数据收集面临挑战

它山之石，可以攻玉



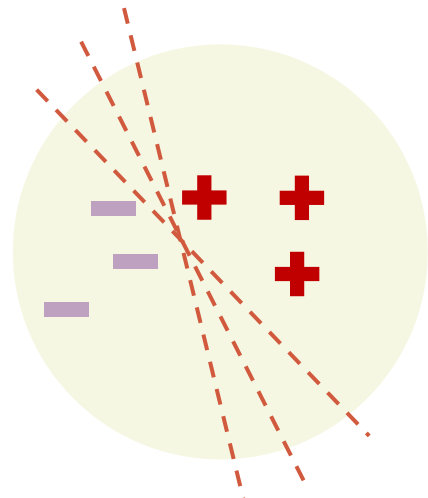
数据收集



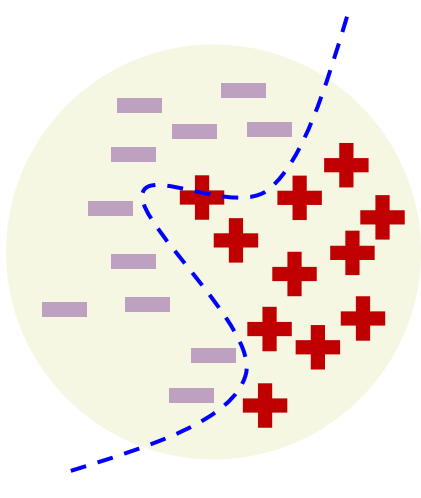
收集代价



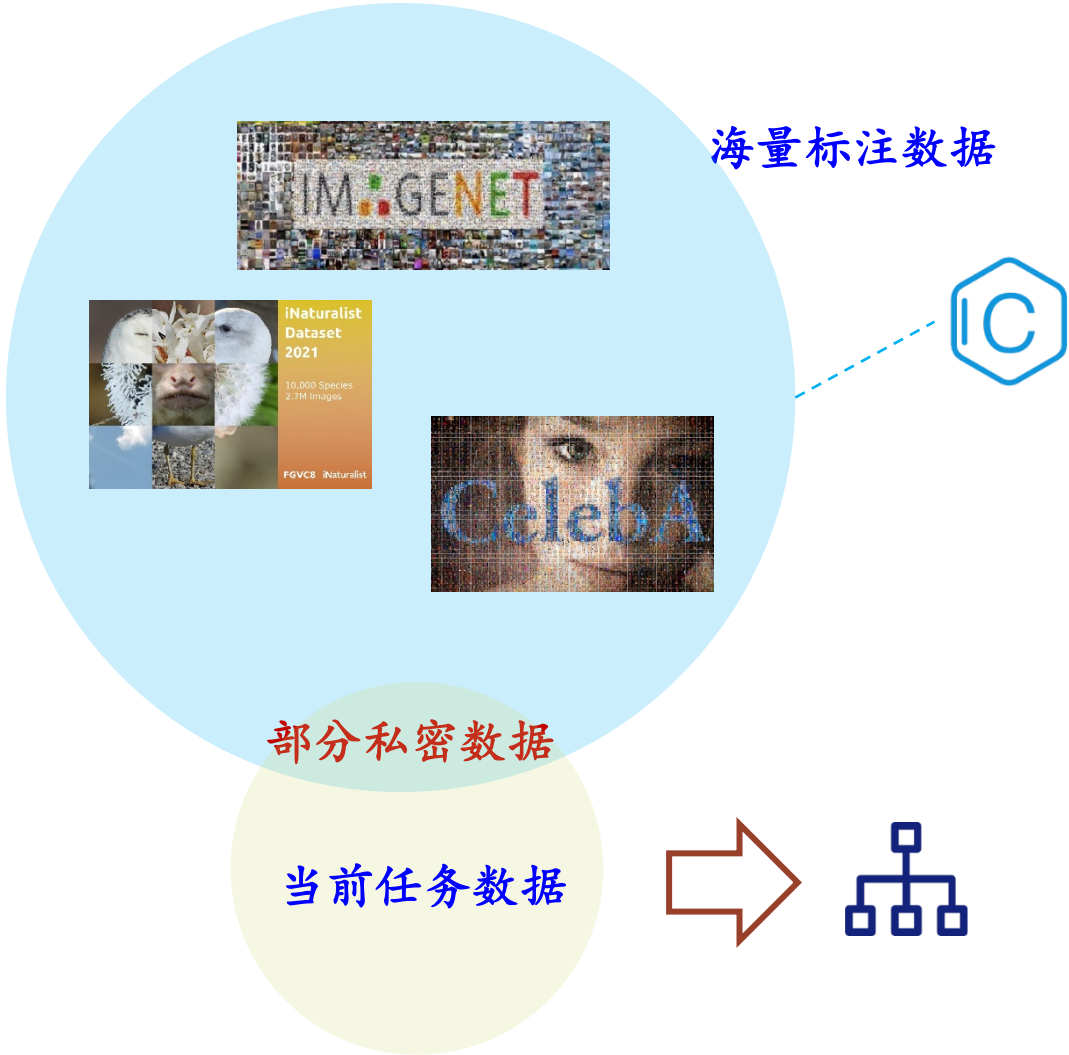
标注代价



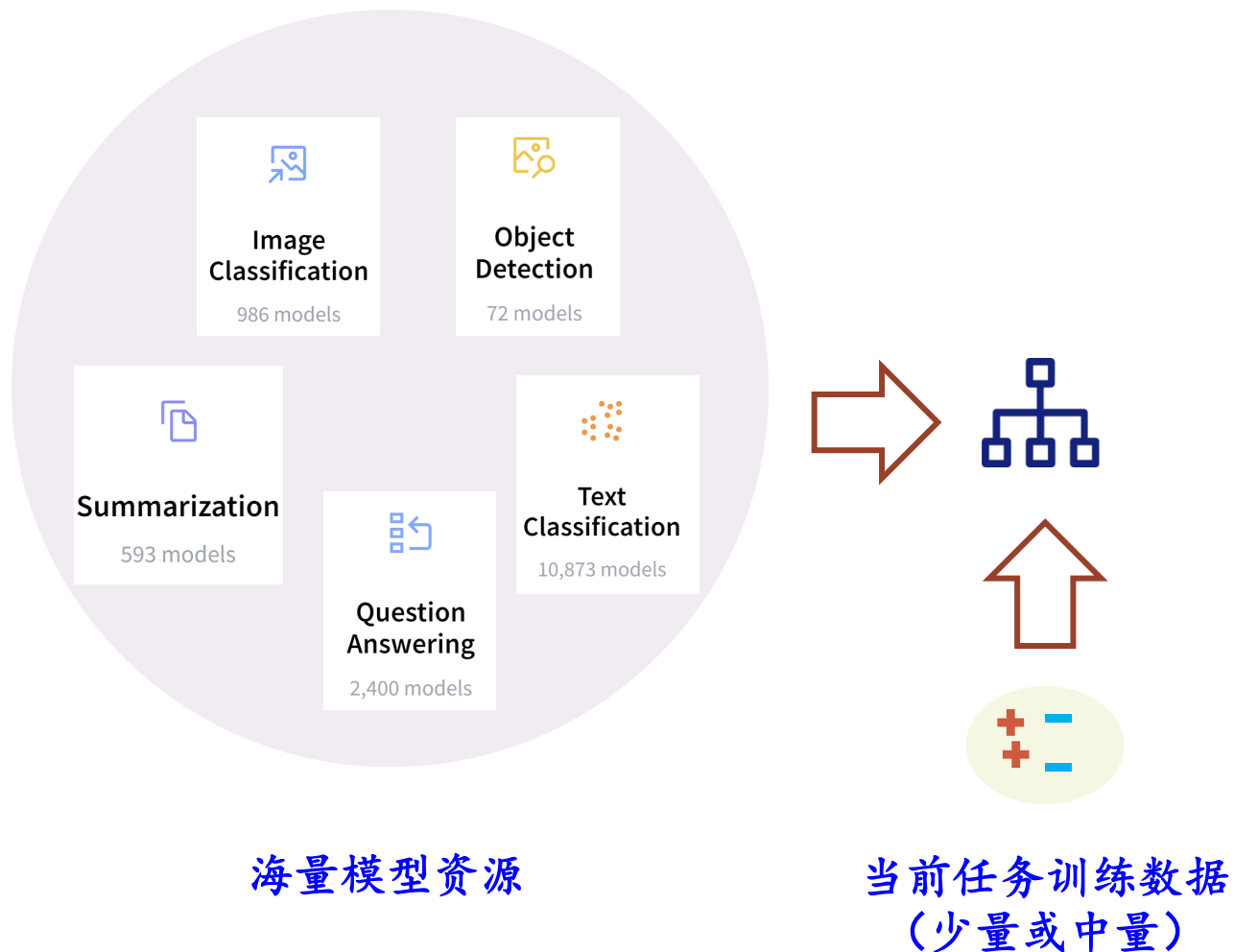
数据不足



数据充分



复用模型



模型复用 (Model Reuse) 问题

给定

(相关任务) 预训练模型 f_{θ_0}

当前任务数据 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$

优化当前模型 f_{θ}

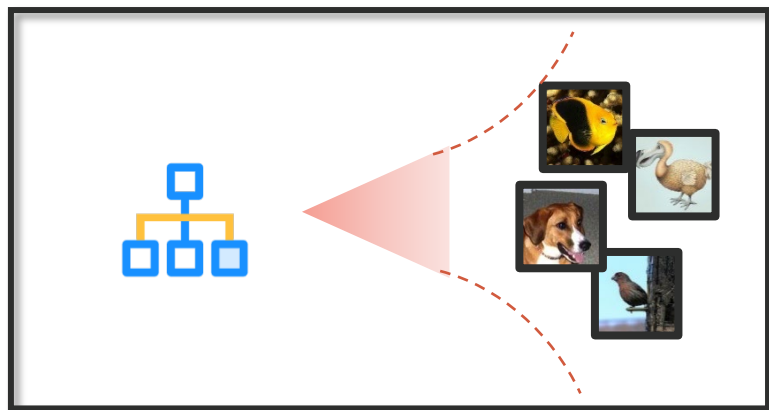
目标

学习 f_{θ} , 优化其泛化性能

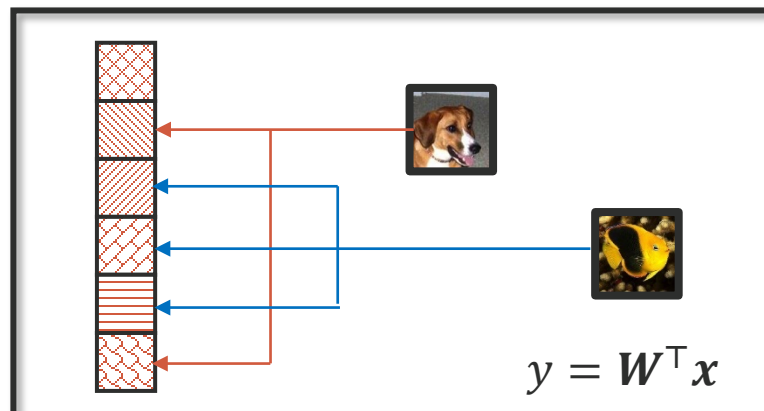
$$E_{(x,y)}[\ell(f_{\theta}(x), y)]$$

模型资源保护原始数据隐私,
同时提供了原始任务的“经验”

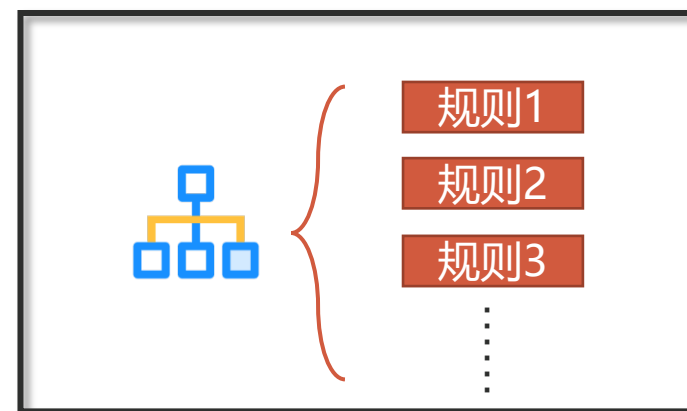
模型所提供的“经验”



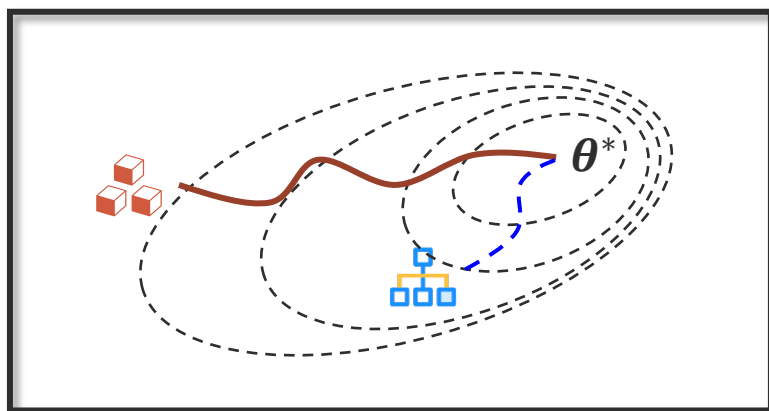
刻画相关数据的分布



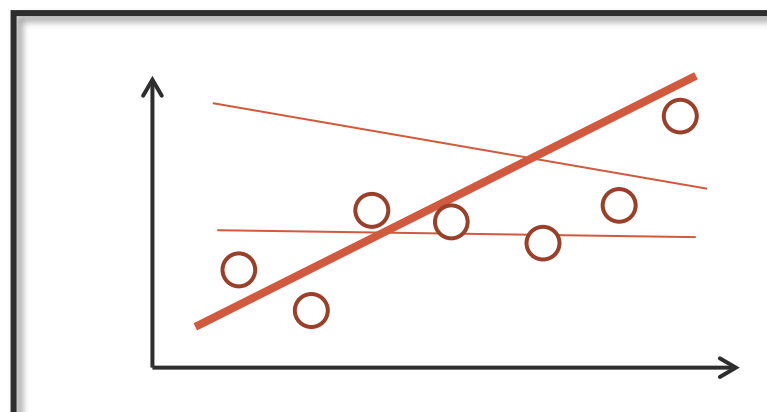
刻画特征和类别如何关联



可解释的专家知识



模型优化的参考 [Kuzborskij and Orabona, ICML'13, MLJ'17]



模型分布的先验 [Rasmussen and Williams, GPML'06][Lam et al., ICML'21]

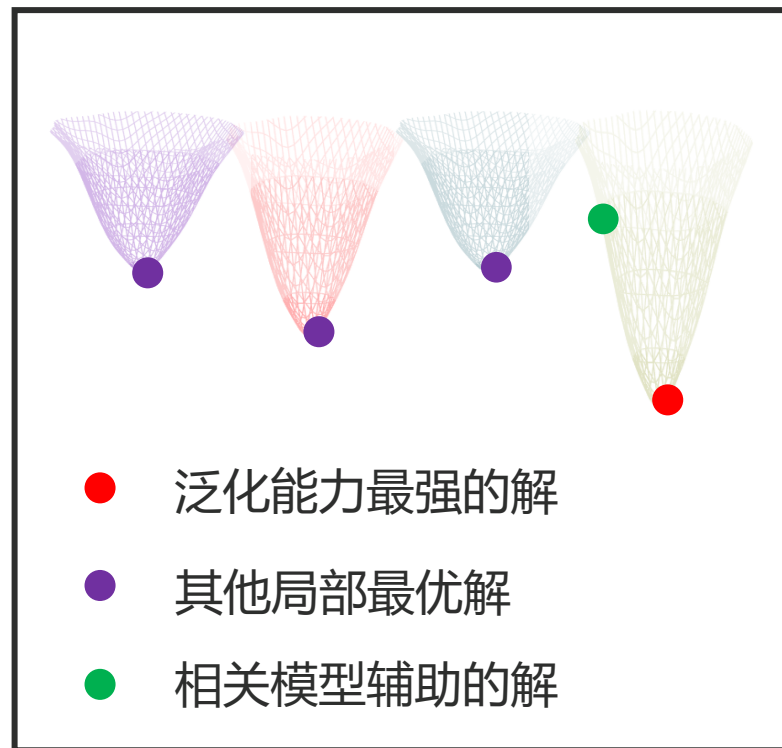
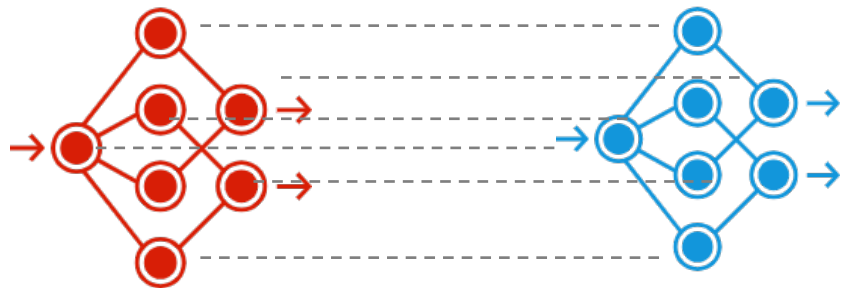
.....

如何利用模型？——参数角度

- 参数匹配：
 - 缩减模型在参数空间中的“距离”
 - 基于参数 θ_0 微调 (Fine-Tune)
 - 参数微调、有偏正则 (Hypothesis Transfer)

[Kuzborskij and Orabona, ICML'13, MLJ'17][Li, Grandvalet, Davoine, ICML'18]

$$\min_{\theta} \ell(f_{\theta}, \mathcal{D}) + \|\theta - \theta_0\|_{\mathcal{H}}$$



复用模型的一些理论启发

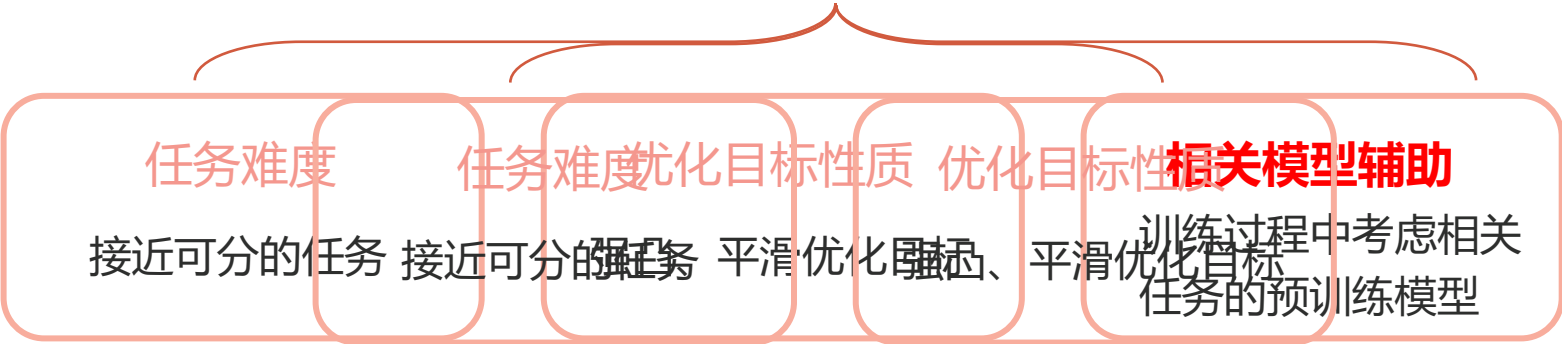
- 模型复用对泛化能力的影响

模型 f_{θ} 的泛化误差

模型 f_{θ} 的训练误差

$$\epsilon(\hat{\theta}) \leq \epsilon_N(\hat{\theta}) + \frac{\text{常数}}{\kappa(N)}$$

如何能够使用更小的样本数 N 得到更紧的泛化界？



一般情况下， $\kappa(N) = \sqrt{N}$ ，能否通过复用模型 f_{θ_0} 获得更紧的泛化界？

复用模型参数辅助当前任务

多分类问题

- 学习 $f_{\theta}(x) = W^{\top} x$
- 相关模型 W_0
- 目标函数

$$\widehat{W} = \operatorname{argmin}_W \sum_{(x_i, y_i)}^N \ell(W^{\top} x_i, y_i) + \lambda \|W - W_0\|_2^2$$

$$\epsilon(\widehat{W}) \leq \epsilon_N(\widehat{W}) + \frac{C_1}{N} + C_2 \sqrt{\frac{\epsilon(W_0)}{N}}$$

更紧的泛化界

与分类器 W_0 的性质有关, W_0 越适配当前任务, 则收敛越快

[Ye, Zhan, Jiang, Zhou, ICML'18][Zhao, Cai, Zhou, MLJ'20] [Ye, Zhan, Jiang, Zhou, TPAMI'21]

度量学习问题

- 学习 $f_{\theta}(x_1, x_2) = \|L^{\top}(x_1 - x_2)\|_2^2$
- 相关模型 L_0
- 目标函数

$$\widehat{L} = \operatorname{argmin}_L \sum_{(x_i, y_i)}^N \sum_{(x_j, y_j)}^N \ell\left(q_{ij} \left(\gamma - \|L^{\top}(x_i - x_j)\|_2^2\right)\right) + \lambda \|L - L_0\|_2^2$$

$$\epsilon(\widehat{L}) \leq \epsilon_N(\widehat{L}) + \frac{C_1}{N} + C_2 \sqrt{\frac{\epsilon(L_0)}{N}}$$

更紧的泛化界

与投影 L_0 的性质有关, L_0 越适配当前任务, 则收敛越快

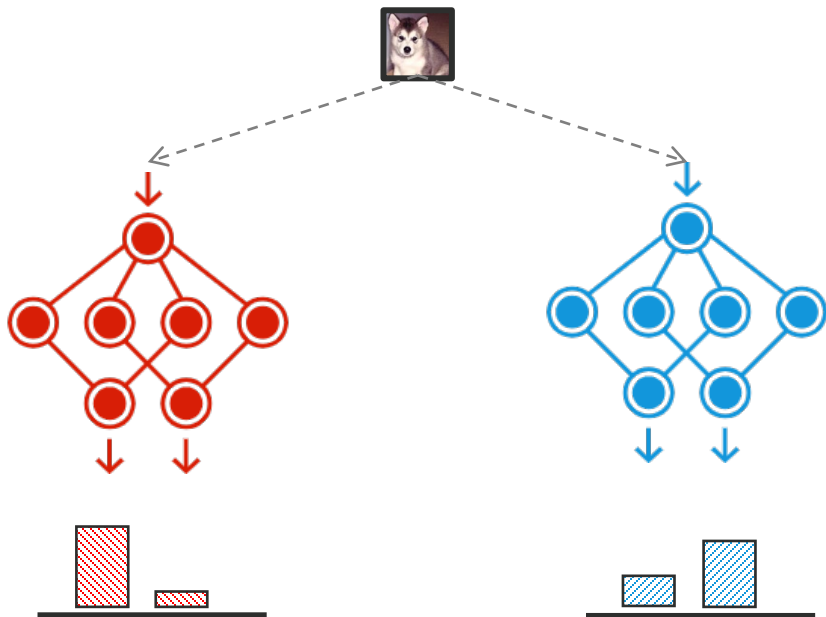
[Ye, Zhan, Li, Jiang, TPAMI'20]

如何利用模型？

数据预测 (Distillation) 角度

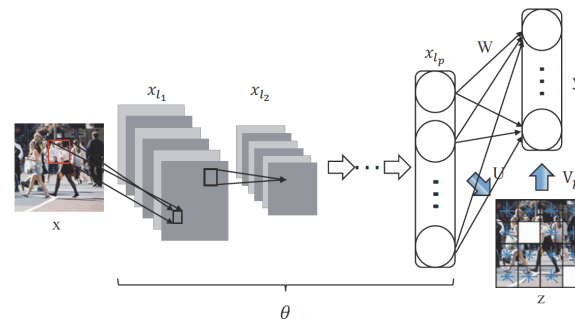
- 缩减模型对数据预测结果的“距离” [Hinton. et al., ArXiv'15][Phuong and Lampert, ICML'19][Gotmare et al., ICLR'19]

$$\min_{\theta} \ell(f_{\theta}, \mathcal{D}) + \|f_{\theta_0}(x) - f_{\theta}(x)\|_{\mathcal{H}}$$



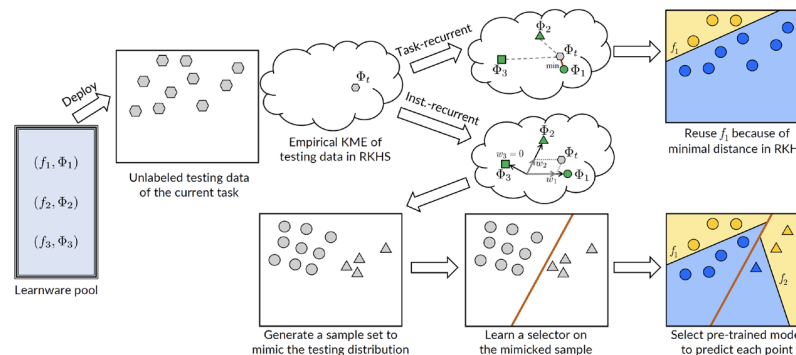
结构化指导

- 通过内部结构实现模型的匹配 [Yang. et al., AAAI'19][Srinivas and Fleuret, ICML'19][Kou et al., NeurIPS'20]



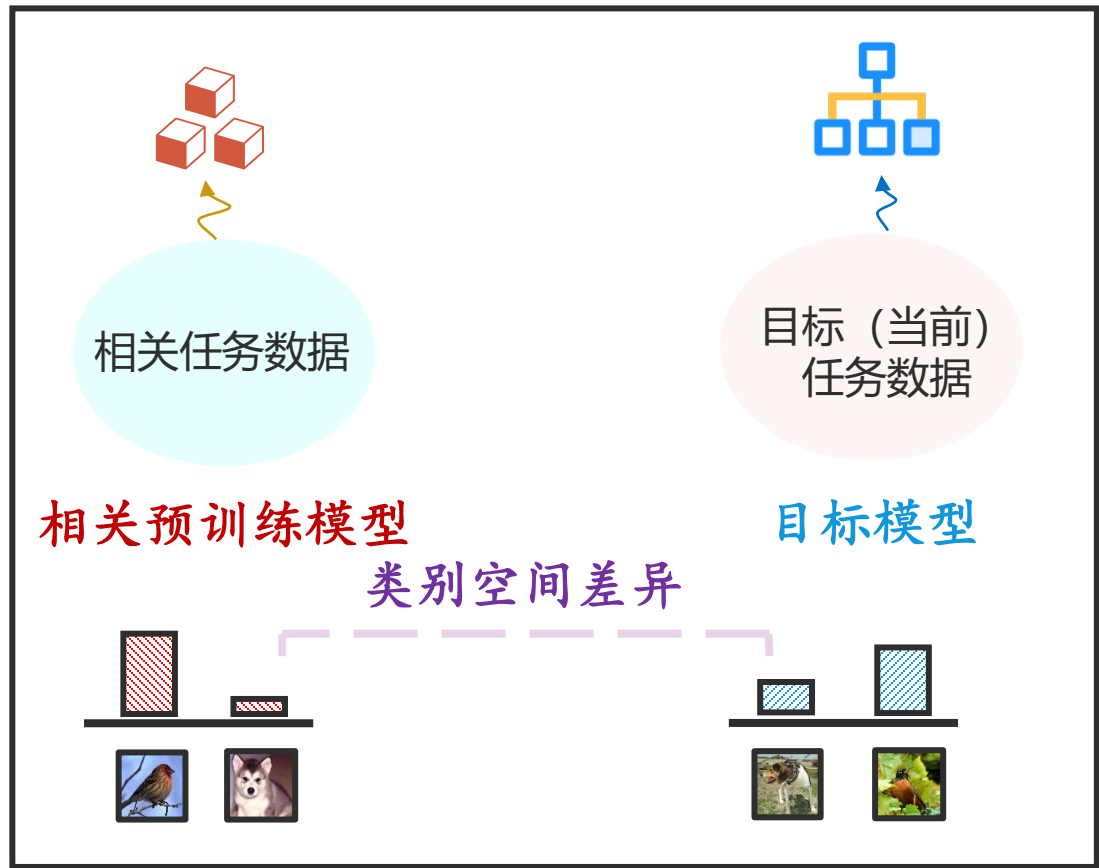
统计量辅助

- 原始模型记录部分统计量，在复用过程中进一步提供帮助 [Wu. et al. TKDE'21]

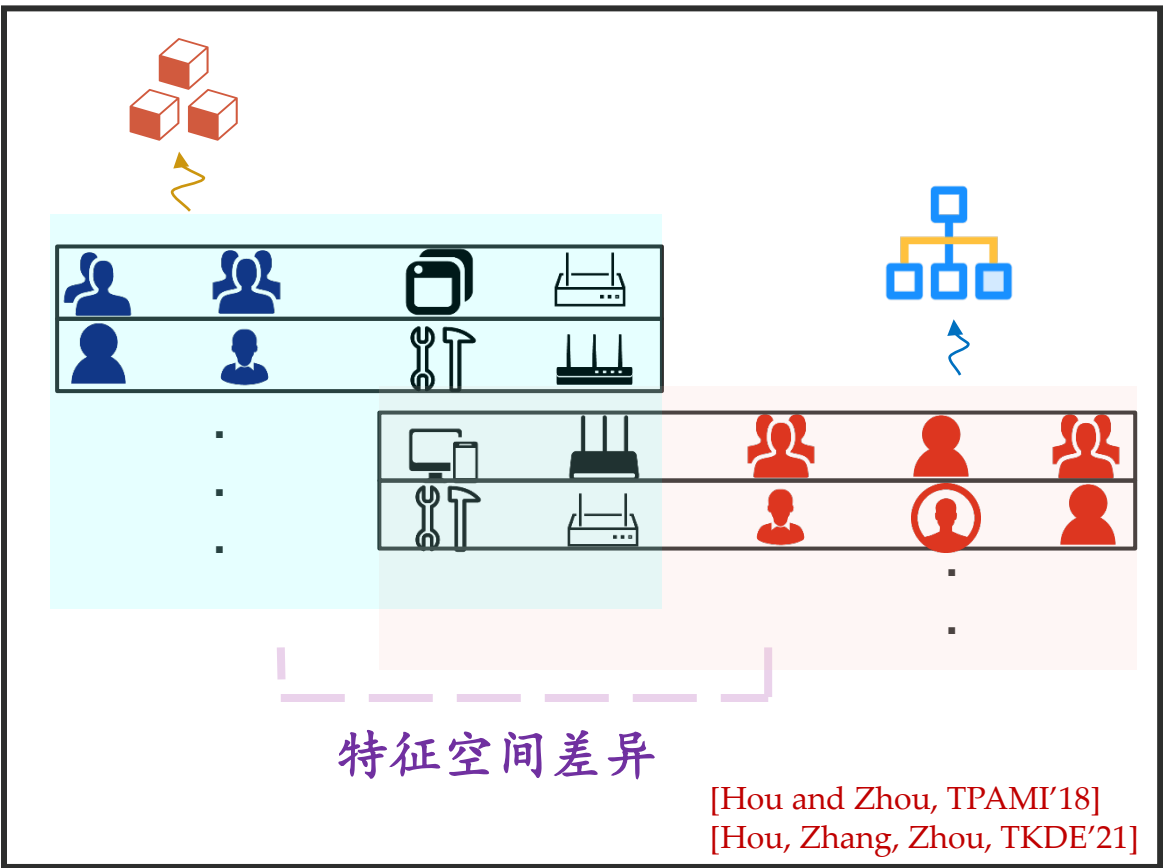


实际任务中的模型复用

相关预训练模型对应的**类别空间**有差异



相关预训练模型对应的**特征空间**有差异



[Hou and Zhou, TPAMI'18]
[Hou, Zhang, Zhou, TKDE'21]

直接“拿来”预训练模型适配行不通

(异构) 模型复用问题

模型复用 (Model Reuse) 问题

给定

(相关任务) 异构预训练模型 g_{θ_0}
当前任务数据 $\{(x_i, y_i)\}_{i=1}^N$
当前模型 f_{θ}

目标

学习 f_{θ} , 优化其泛化性能

$$E_{(x,y)}[\ell(f_{\theta}(x), y)]$$

从 “雪中送炭” 到 “锦上添花”

参数对齐

$$\min_{\theta} \ell(f_{\theta}, \mathcal{D}) + \|\theta_0 - \theta\|_{\mathcal{H}}$$

相关任务模型 g_{θ_0} 和当前模型 f_{θ} 不匹配, 无法直接参数对齐 [Singh and Jaggi, NeurIPS'20]

预测对齐

$$\min_{\theta} \ell(f_{\theta}, \mathcal{D}) + \|g_{\theta_0}(x) - f_{\theta}(x)\|_{\mathcal{H}}$$

相关任务模型 g_{θ_0} 和当前模型 f_{θ} 的预测结果维度不一致, 无法直接对齐

异构场景中常见的复用方法

Fine-Tune

```

$$\theta' = \text{InLoop}(\mathcal{D}, \theta_0, M)$$

$$\theta' \leftarrow \theta_0$$

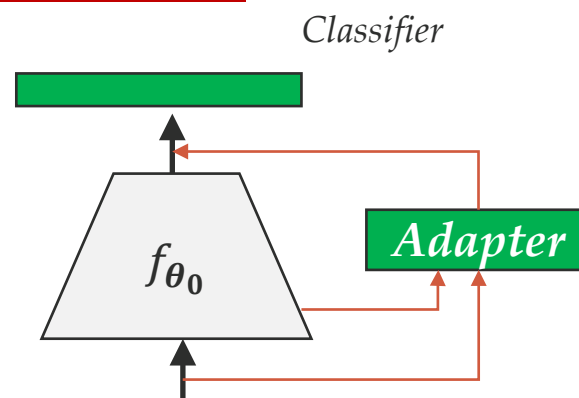
$$\text{for } m \in [M] \text{ do}$$

$$\theta' = \theta' - \alpha \nabla_{\theta'} \mathcal{L}(\mathcal{D}, \theta')$$

```

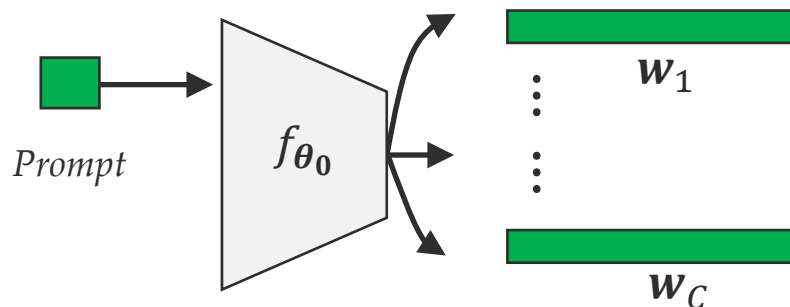
对原始模型部分参数进行调节

Adaptation



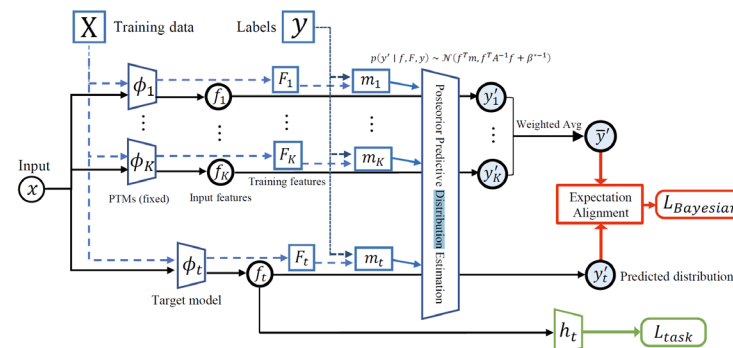
增加中间层结构进行适配[He et al., ICLR'22]

Prompt Tuning



增加少量可学参数作为提示[Bahng et al., NeurIPS'22][Jia et al., ECCV'22][Zhou et al., IJCV'22]

B-Tuning

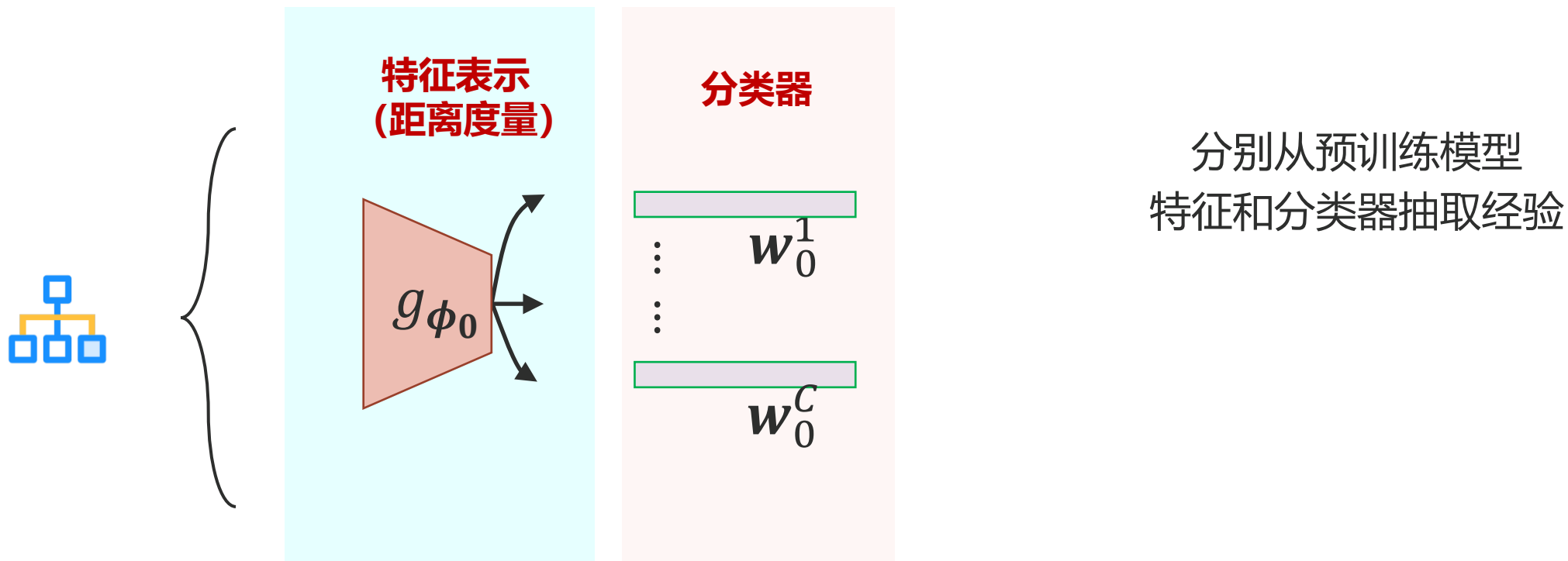


后验分布估计并匹配[You et al., JMLR'22]

模型复用：一种解耦的视角

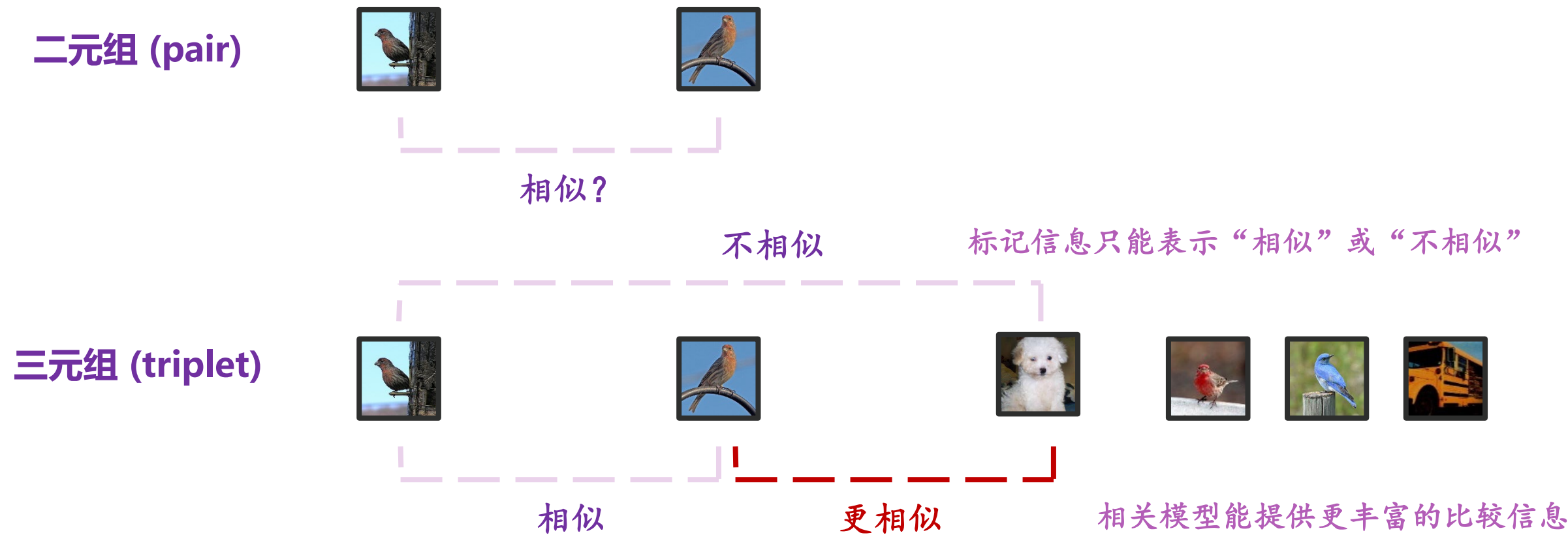
- 从分类器和特征两个方面考虑复用问题

基于相关任务数据的相似度让相似的样本近，
不相似的远（可以跨任务迁移）



描述预训练模型特征和原始类别之间的
对应关系（难以跨任务迁移）

模型对于相似关系的匹配能力

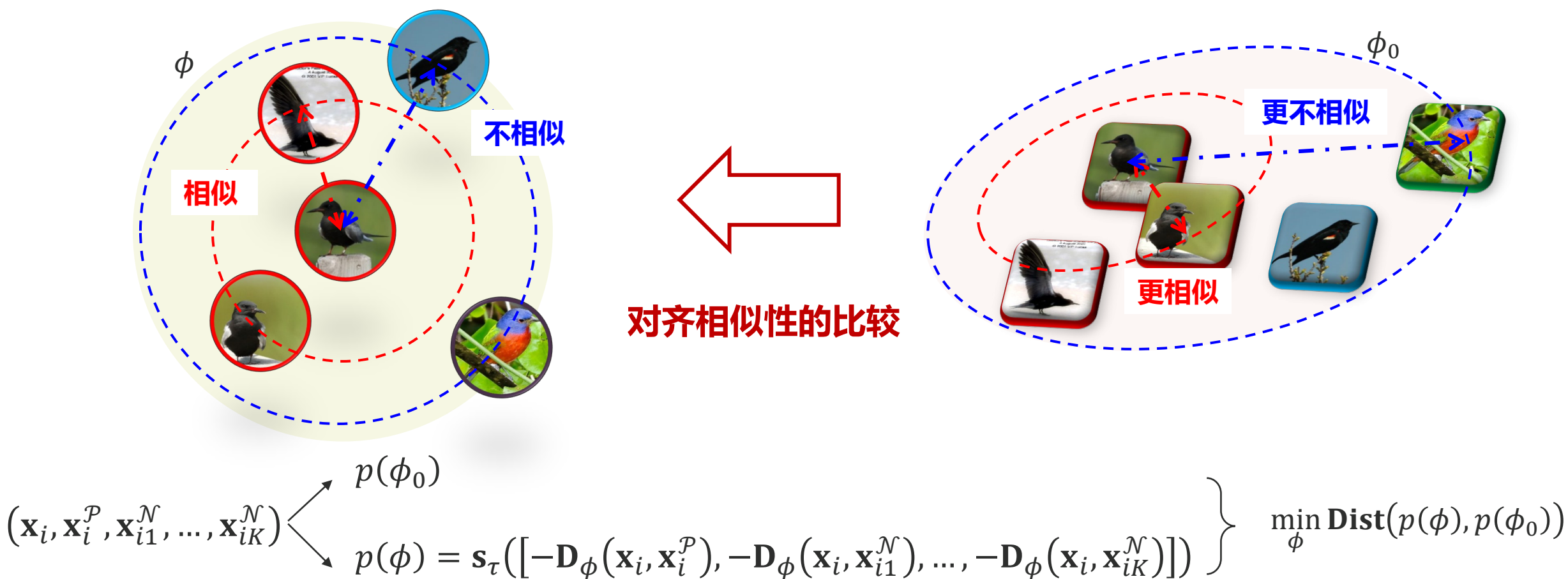


通过模仿相似关系的判断能力，为当前任务特征表示的学习提供经验

模型之间表示能力的迁移

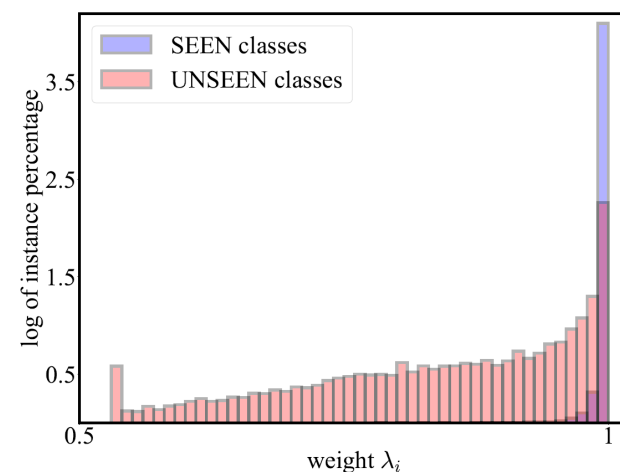
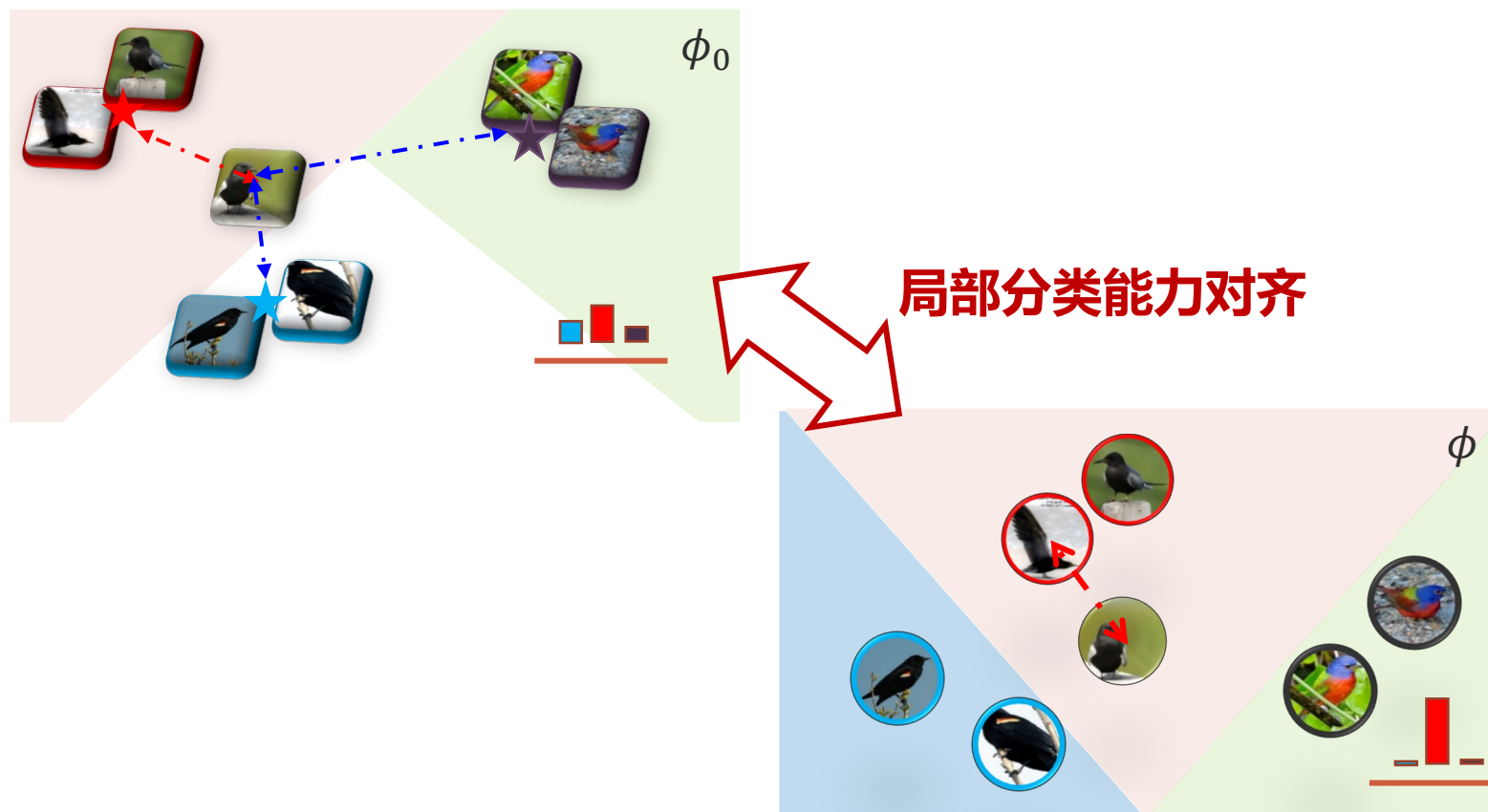
REFILLED: 使当前任务的特征表示 ϕ 能够有效“模仿”已有模型的特征表示 ϕ_0

- **对齐**当前任务特征和目标模型特征下的相似关系比较的概率



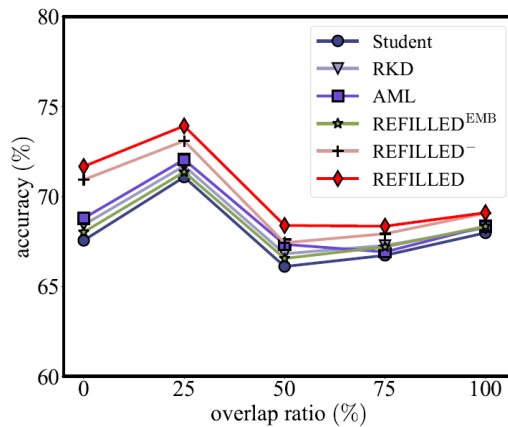
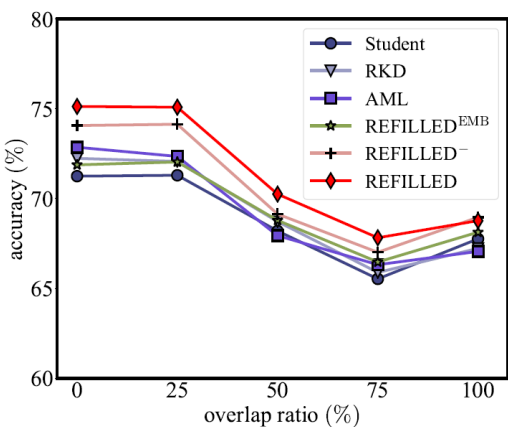
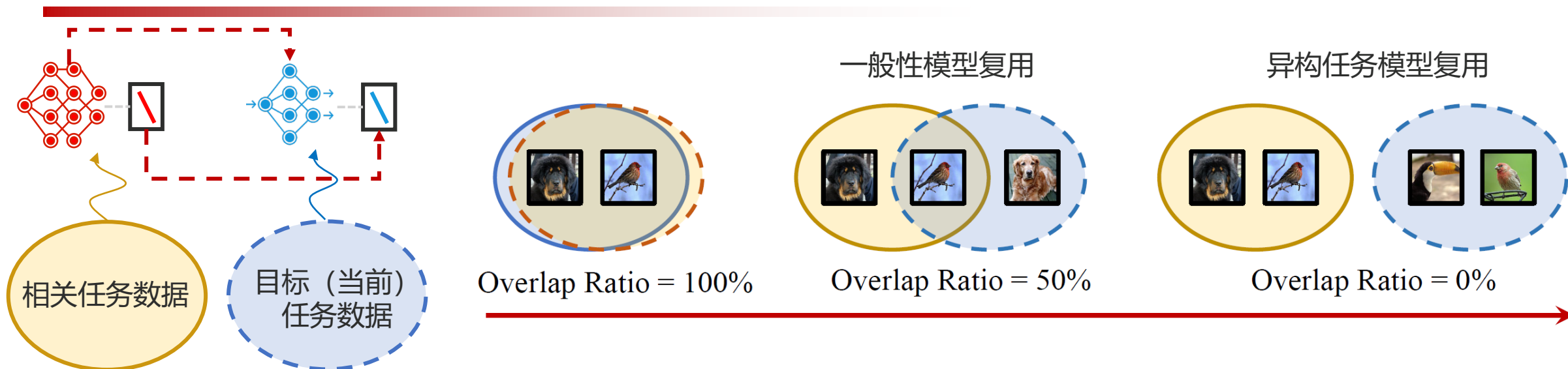
模型之间局部分类能力的迁移

分类器复用：利用已有模型的特征表示 ϕ_0 构造近邻分类器，指导当前任务的模型学习，提供辅助监督信息

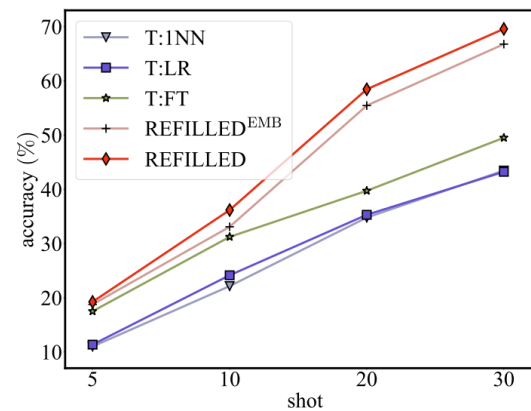


基于置信度“选择性”指导

异构类别模型复用实验



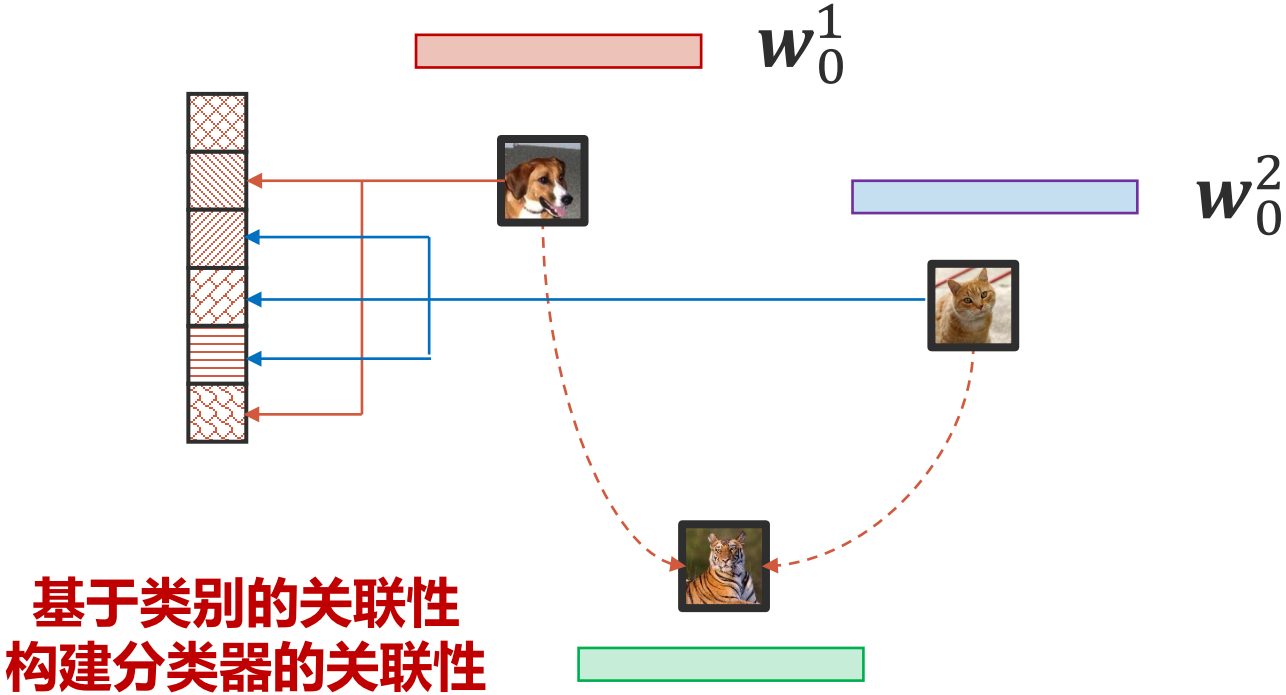
f_{θ} 和 g_{θ_0} 相同结构 (MobileNets-1.0, 左)
或不同结构 (MobileNets-0.75, 右)



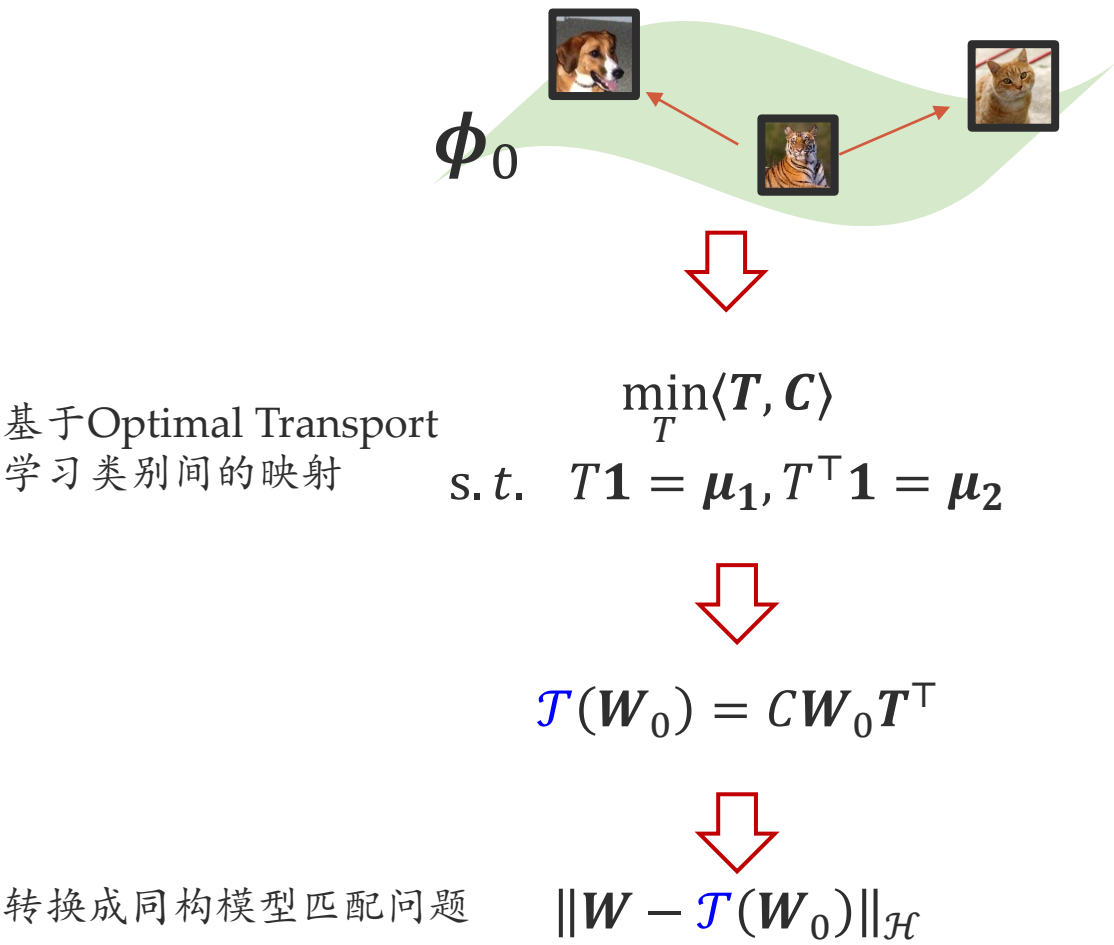
不同样本数量下复用

从语义映射到分类器复用

- 利用类别的关联性，将原始任务的分类器转化成应对当前任务的分类器



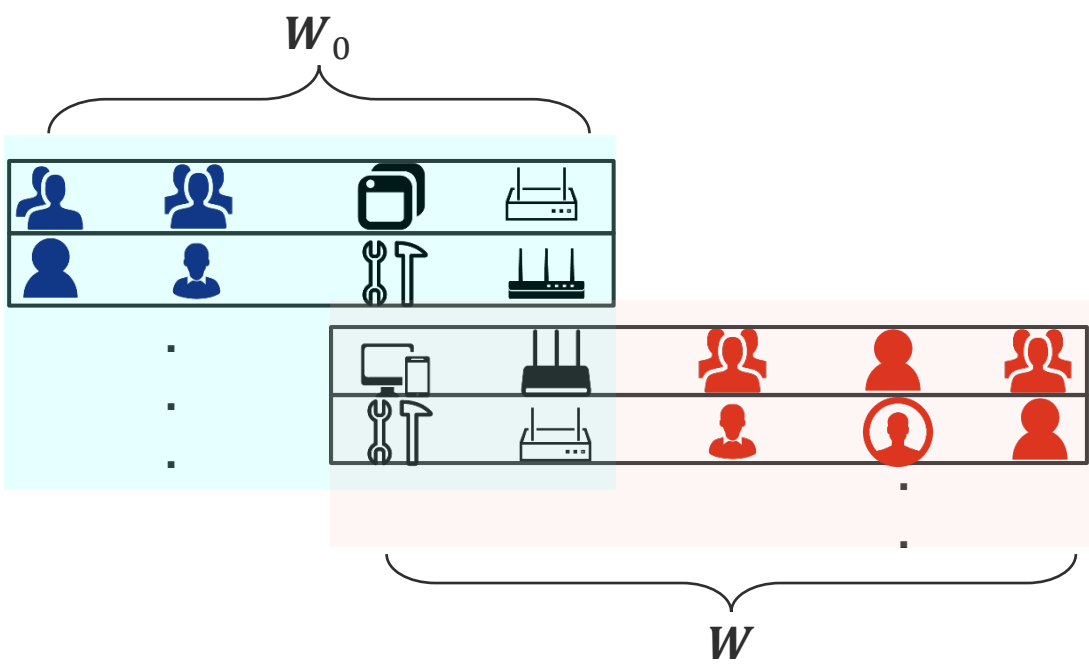
$$\begin{aligned} \text{[Tiger Image]} &= \alpha \text{ [Dog Image]} + \beta \text{ [Cat Image]} \\ \text{[Green Bar]} &= \alpha \text{ [Red Bar]} + \beta \text{ [Blue Bar]} \end{aligned}$$



异构特征的模型复用

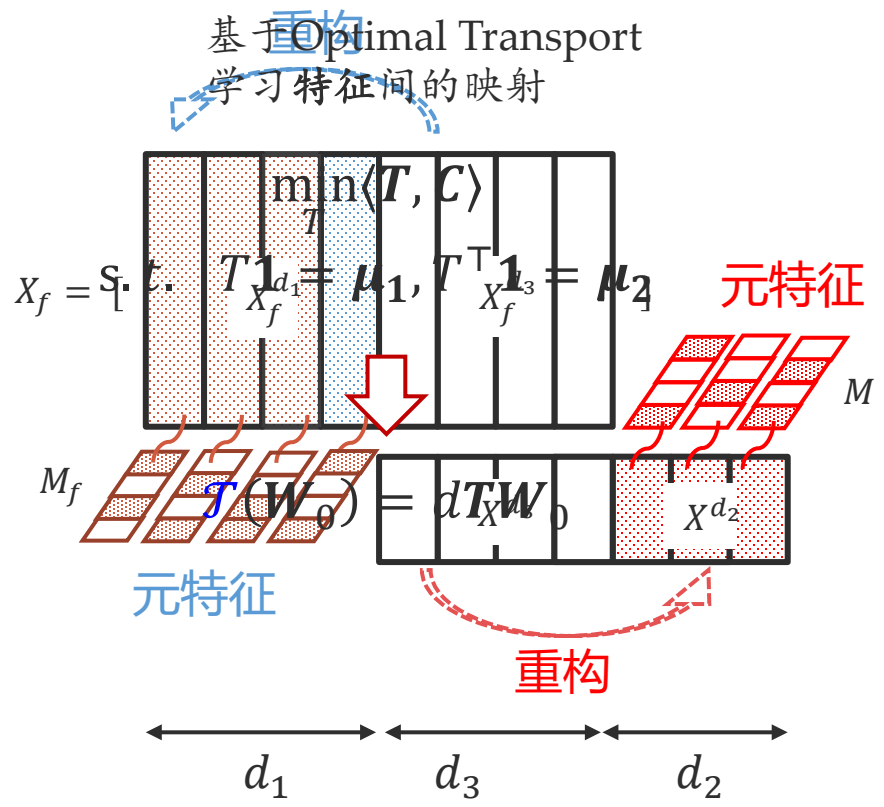
ReForm: 在两个分类器之间学习语义映射 $\mathcal{T}$:

$$\min_{\theta} \sum_{i=1}^N \ell(f_{\theta}(x_i), y_i) + \lambda \mathcal{R}(\mathcal{T}(W_0), W)$$



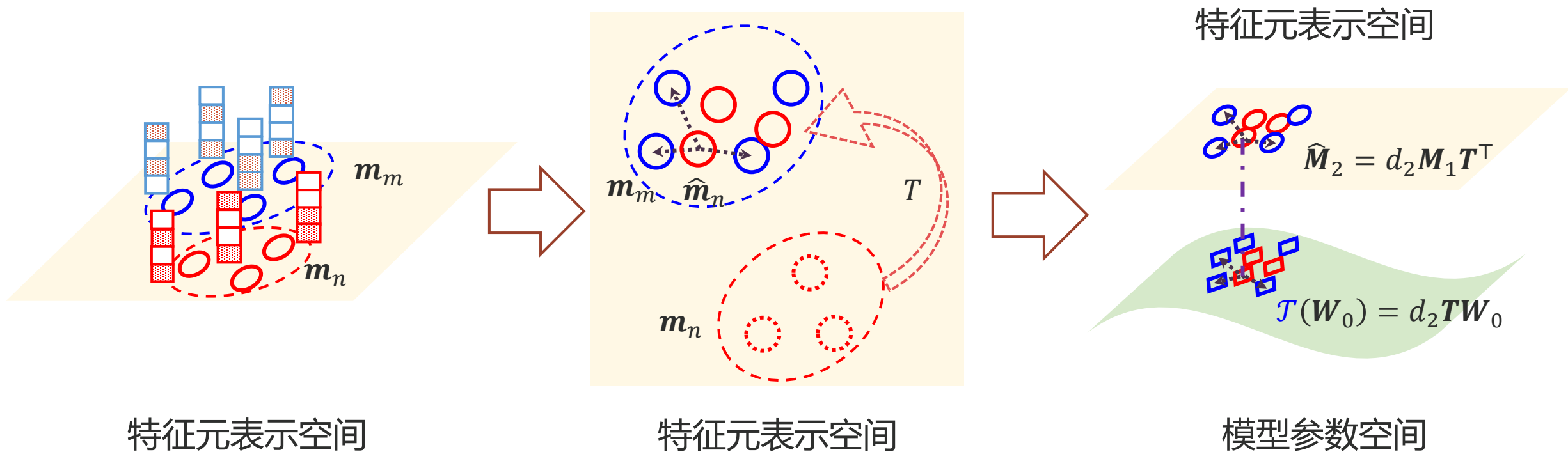
特征空间差异

对每一维特征产生 “元表示 (Meta Representation)”

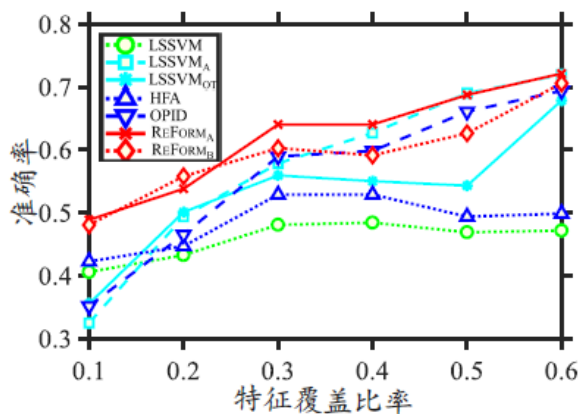


异构特征的模型复用

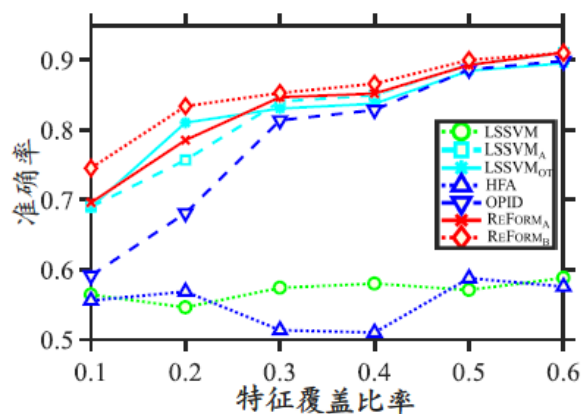
在特征元空间中利用Optimal Transport获得原始特征和新特征之间的关系



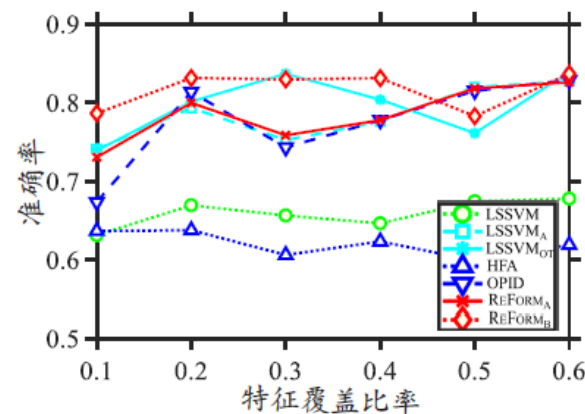
异构特征的模型复用



(a) mfeat_fou

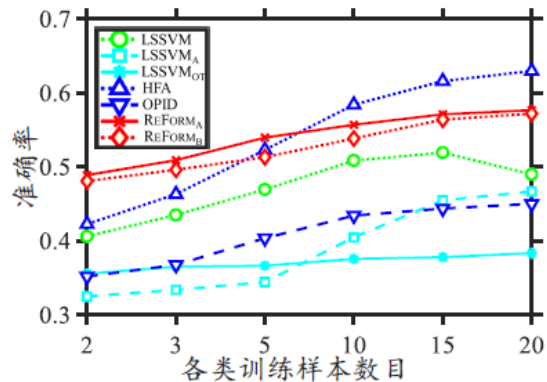


(c) reut8

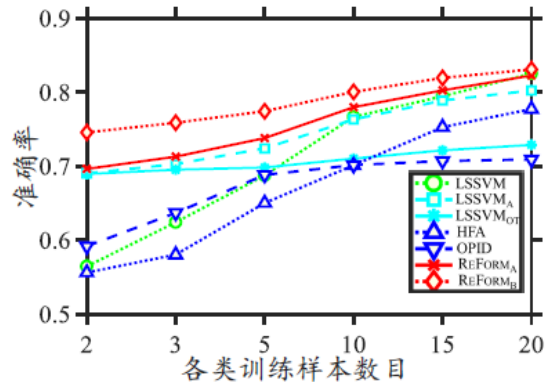


(e) spambase

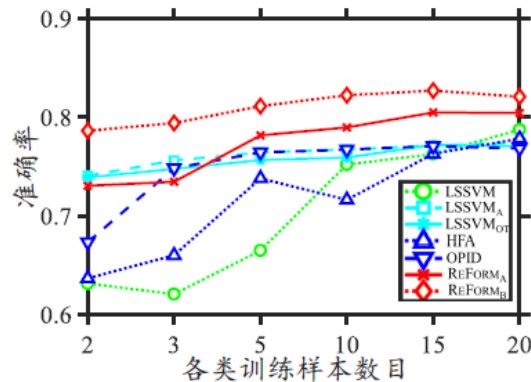
任务之间特征的覆盖比率增大



(b) mfeat_fou



(d) reut8

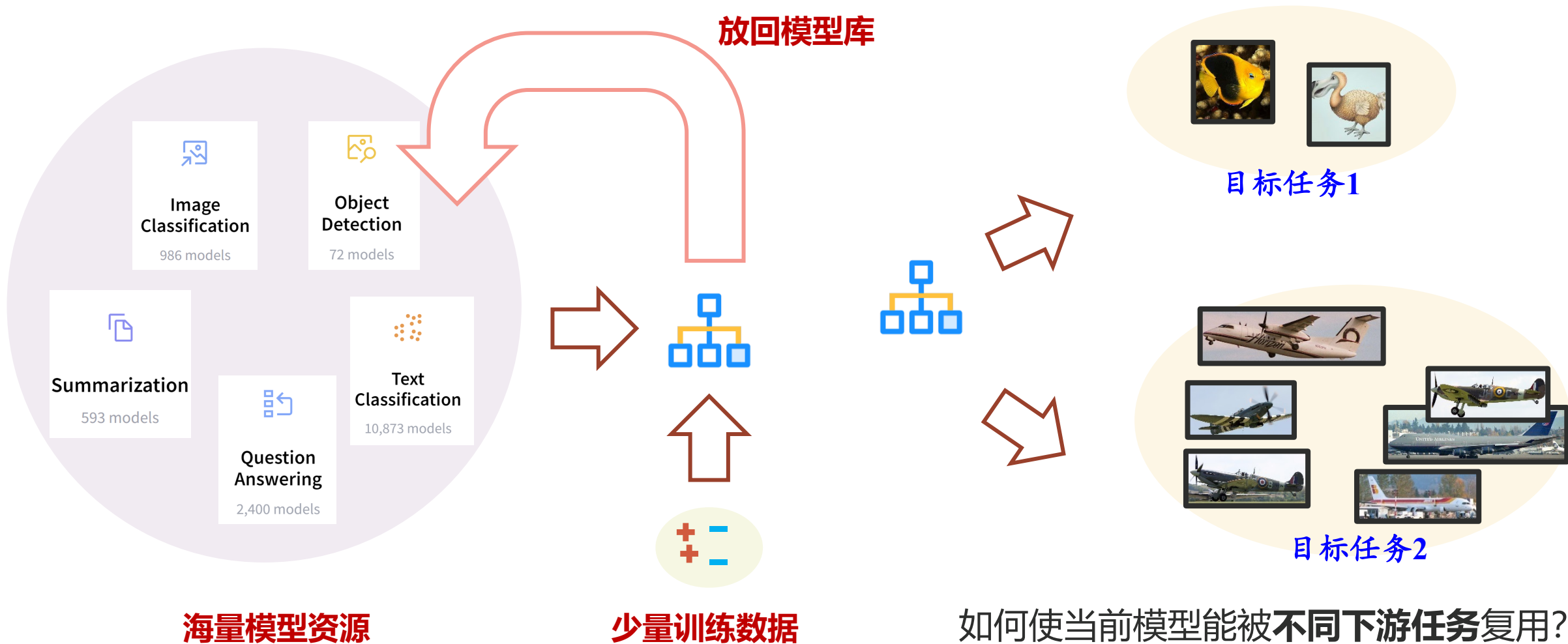


(f) spambase

当前任务训练样本数目增多

模型可复用：输出当前模型能力

- 当前学到的模型可能会在未来场景中被其他模型复用



学习可复用的模型

可复用模型 (Reusable Model)

给定

数据 $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{N_0}$

学习模型 g_{θ_0}

目标

给定新任务 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$

$$f_{\theta} = \mathcal{A}(g_{\theta_0}, \mathcal{D})$$

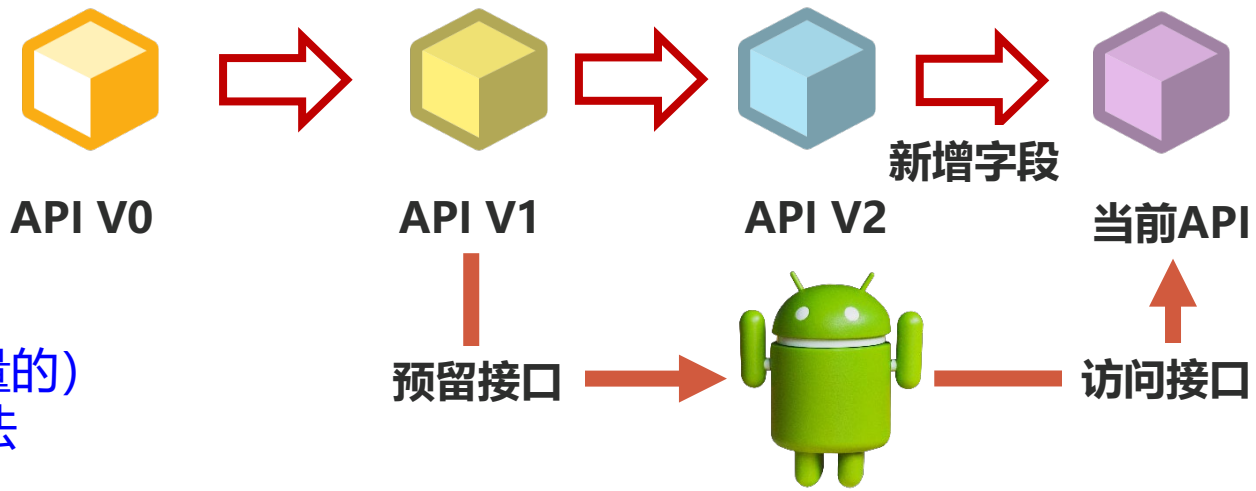
具有较小的泛化误差

$$E_{(x,y) \sim \mathcal{D}}[\ell(f_{\theta}(x, y))]$$

一种 (轻量的)
复用方法

使预训练的模型 “更好用”

为后续应用 “埋下伏笔”



可复用模型的形式与难点

$\mathcal{A}$:特征表示

基于 ϕ_0 计算 $\mathcal{D}$ 的特征, 进行近邻分类 [Vinyals et al., NIPS'16][Snell, Swersky, Zemel, NIPS'17]

$$\mathcal{A}(g_{\theta_0}, \mathcal{D}) = \text{NCM}(\phi_0(\mathcal{D}))$$

$\mathcal{A}$:优化方法

基于 θ_0 在 $\mathcal{D}$ 上进行一定步数的Fine-Tune [Finn, Abbeel, Levine, ICML'17][Nichol, Achiam, Schulman, ArXiv'18]

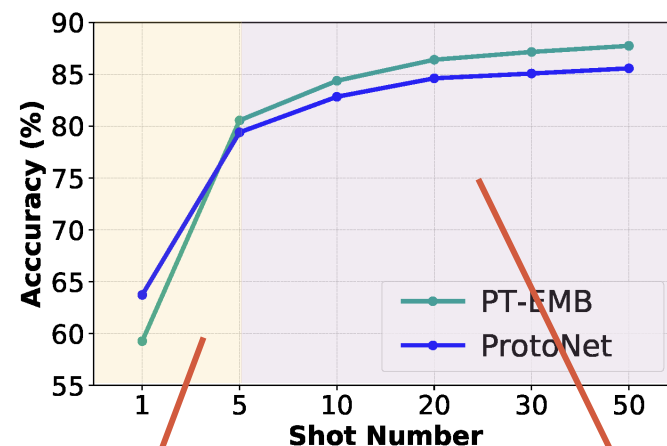
$$\mathcal{A}(g_{\theta_0}, \mathcal{D}) = \text{InLoop}(\mathcal{D}, \theta_0, M)$$

$\mathcal{A}$:线性模型

基于 ϕ_0 在 $\mathcal{D}$ 上训练线性模型 [Lee et al., CVPR'19][Raghu et al., ICLR'20]

$$\mathcal{A}(g_{\theta_0}, \mathcal{D}) = \text{SVM}(\phi_0(\mathcal{D}))$$

⋮



样本少 →
特殊复用方法调整的特征
表示更优

[Vinyals et al., NIPS'16]
[Snell, Swersky, Zemel, NIPS'17]

样本较多 →
直接复用预训练模型的
特征表示更优

[Wang et al., ArXiv'19]
[Tian et al., ECCV'20]

学习可复用的分类器

监督学习

学习 样本 $\rightarrow$ 标记 的映射

(, 车) (, 船)

⋮

(, 狗) (, 海)

(, ?)

(, ?)

(, ?)

元学习 (Meta-Learning)

学习 目标任务 $\rightarrow$ 对应目标模型 的映射

已知数据
(预训练)

(  , $f_{\theta_0^1}^*$)

训练集1

验证集1

(  , $f_{\theta_0^2}^*$)

训练集2

验证集2

(  , ?)

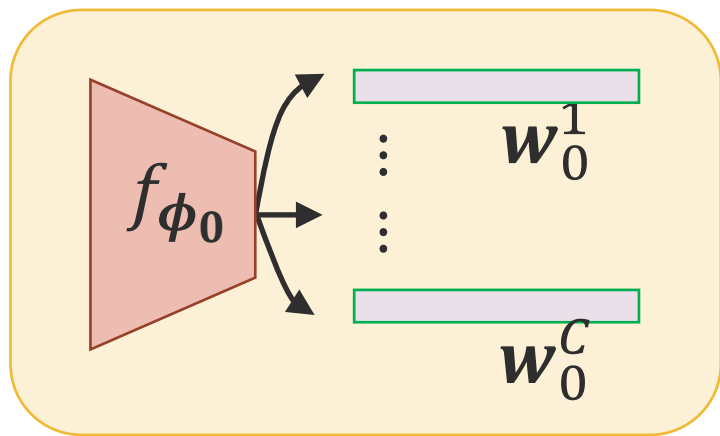
训练集D

目标任务数据

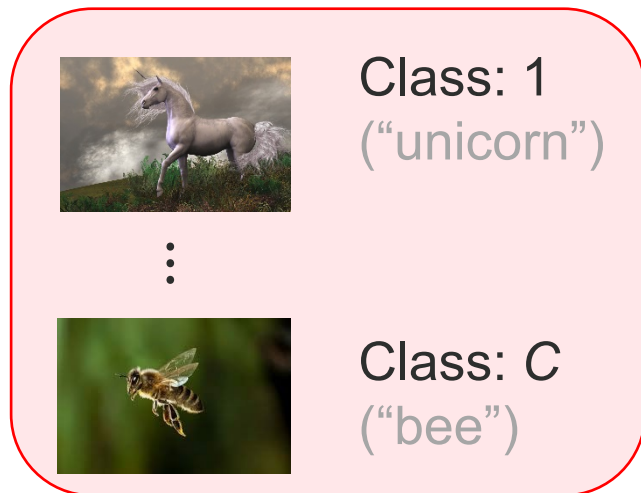
$\mathcal{A}$

学习可复用的分类器

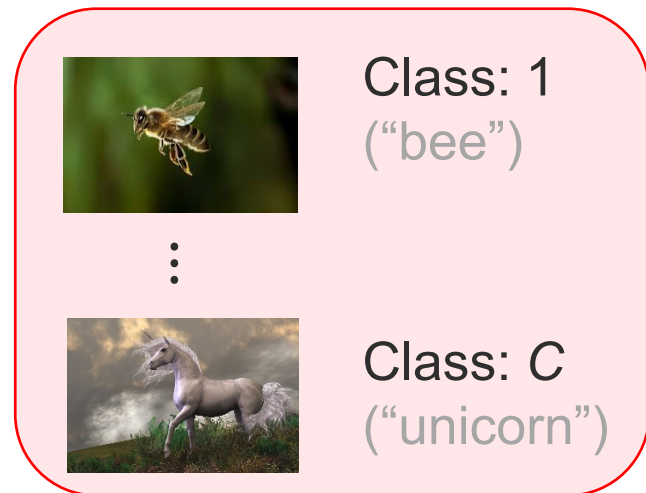
模型初始化



(下游任务) 小样本训练集



(下游任务) 小样本训练集2



...

类别顺序置换

优化步骤

```

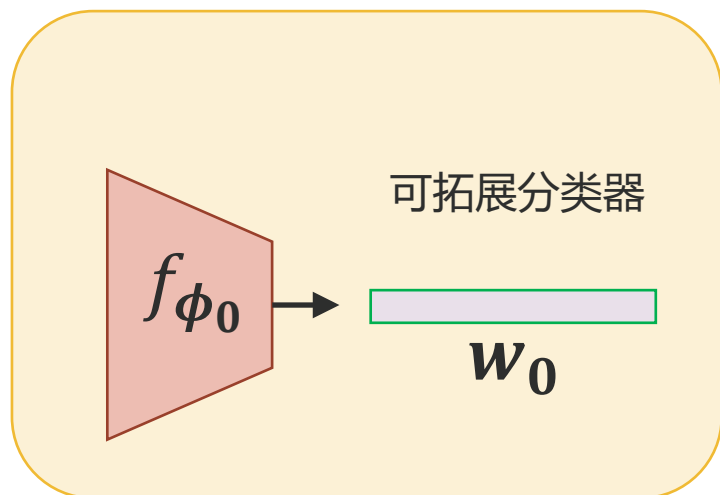
$$\begin{aligned} \theta' &= \text{InLoop}(\mathcal{D}, \theta_0, M) \\ \theta' &\leftarrow \theta_0 \\ \text{for } m \in [M] \text{ do} \\ \theta' &= \theta' - \alpha \nabla_{\theta'} \mathcal{L}(\mathcal{D}, \theta') \end{aligned}$$

```

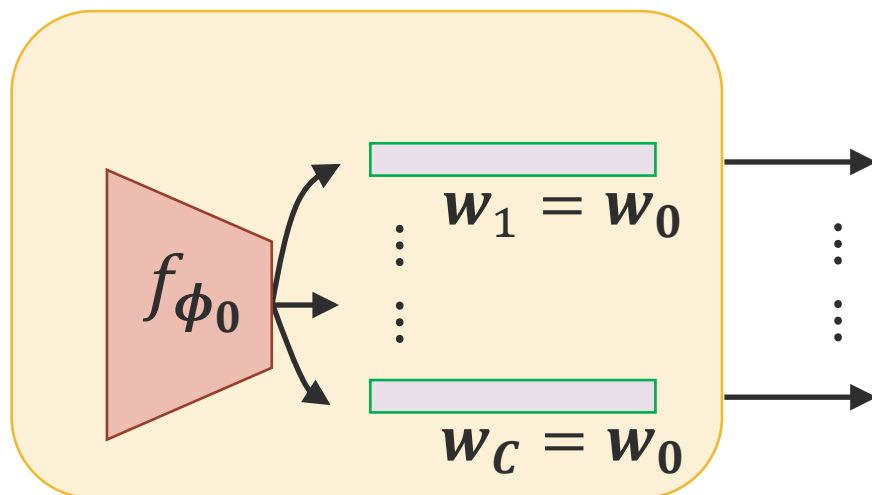
学习可复用的分类器

Unicorn-MAML: 预训练阶段学习对所有类别通用的（可拓展）分类器，方便在下游任务中方便复用

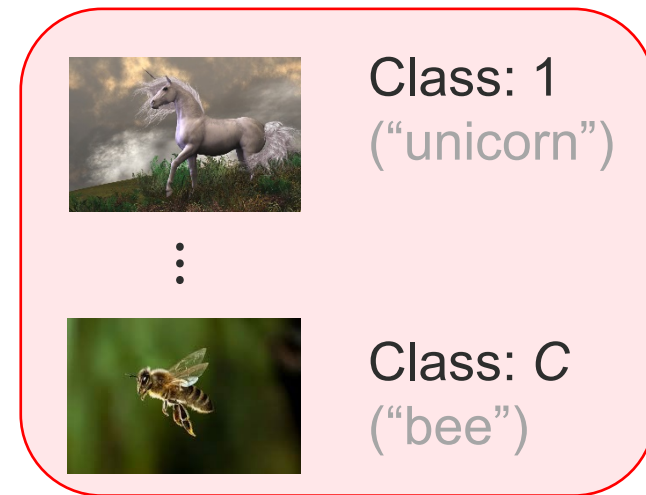
预训练模型构造可复用参数
(meta-parameters θ_0)



分类器拓展



拓展后的分类器
在目标任务数据进行优化



学习可复用的分类器

Multi-Shot	10-shot	20-shot	30-shot	50-shot
SimpleShot (Wang et al., 2019)	84.89	86.91	87.53	88.08
ProtoNet (Snell et al., 2017)	82.83	84.61	85.07	85.57
FEAT (Ye et al., 2020a)	85.15	87.09	87.82	87.83
MAML (our implementation)	88.08	90.23	91.06	92.14
UNICORN-MAML	88.38	90.96	91.96	92.86

在MiniImageNet数据集尝试每类10/20/30/50分类任务

MiniImageNet → CUB	1-Shot	5-Shot
Baseline++ (Chen et al., 2019)	50.37	73.30
ProtoNet (Snell et al., 2017)	50.01	72.02
Neg-Cosine (Liu et al., 2020)	47.74	69.30
MAML (our implementation)	51.25	73.86
UNICORN-MAML	51.80	75.67

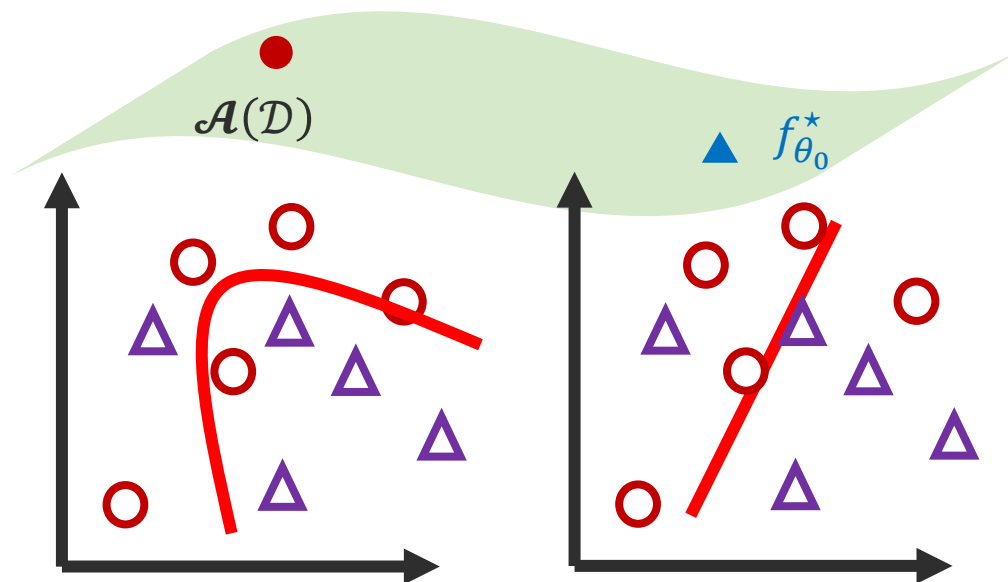
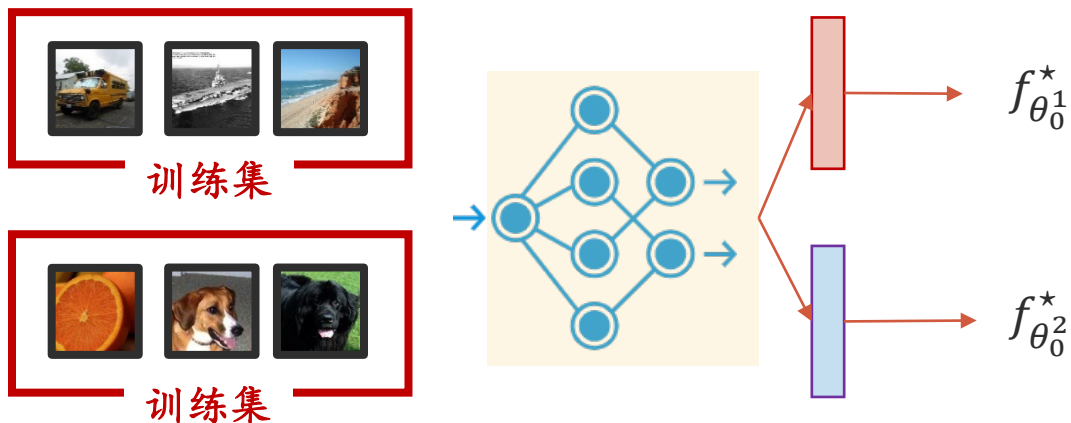
在MiniImageNet数据集学习可复用模型，
应用于CUB数据集每类1/5样本的分类任务

构架复用性更强的模型作为指导

学习 目标任务 $\rightarrow$ 对应目标模型 的映射



LastShot方法提升分类器、特征的可复用性

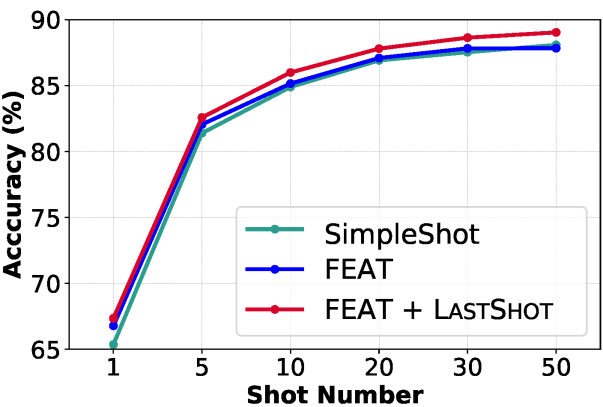
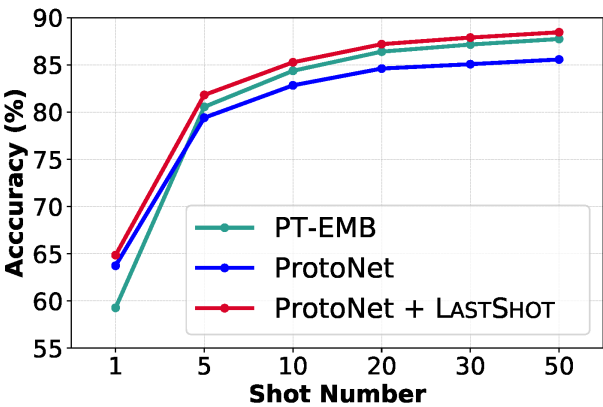


LastShot实验结果

可作为通用的框架应用于多种小样本下游任务中，产生提升

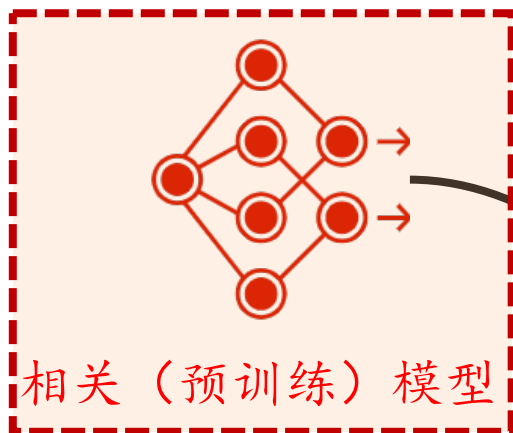
Setting	1-Shot	5-Shot
ModelRegression [16]	61.94 ± 0.20	76.24 ± 0.14
MetaOptNet [25]	62.64 ± 0.20	78.63 ± 0.14
SimpleShot [†] [9]	65.36 ± 0.20	81.39 ± 0.14
TRAML+ProtoNet [107]	60.31 ± 0.48	77.94 ± 0.57
RFS-Simple [34]	62.02 ± 0.63	79.64 ± 0.44
RFS-Distill [34]	64.82 ± 0.60	82.14 ± 0.43
DSN-MR [108]	64.60 ± 0.72	79.51 ± 0.50
MTL+E3BM [109]	63.80 ± 0.40	80.10 ± 0.30
DeepEMD [35]	65.91 ± 0.82	82.41 ± 0.56
TRAML+AM3 [107]	67.10 ± 0.54	79.54 ± 0.60
ProtoNet [†] [13]	62.39 ± 0.20	79.74 ± 0.14
+ LASTSHOT (NC)	64.16 ± 0.20	81.23 ± 0.14
+ LASTSHOT (LR)	64.80 ± 0.20	81.65 ± 0.14
ProtoMAML [†] [24]	62.04 ± 0.21	79.62 ± 0.14
+ LASTSHOT (NC)	63.07 ± 0.20	81.04 ± 0.14
+ LASTSHOT (LR)	63.05 ± 0.20	81.11 ± 0.14
MetaOptNet [†] [25]	63.21 ± 0.20	79.94 ± 0.14
+ LASTSHOT (NC)	65.07 ± 0.20	80.34 ± 0.14
+ LASTSHOT (LR)	65.08 ± 0.20	80.40 ± 0.14
FEAT [23]	66.78 ± 0.20	82.05 ± 0.14
+ LASTSHOT (NC)	67.33 ± 0.20	82.39 ± 0.14
+ LASTSHOT (LR)	67.35 ± 0.20	82.58 ± 0.14

Setups →	1	5	10	20	30	50
PT-EMB	59.27 ± 0.20	80.55 ± 0.14	84.37 ± 0.12	86.40 ± 0.11	87.15 ± 0.10	87.74 ± 0.10
SimpleShot [9]	65.36 ± 0.20	81.39 ± 0.14	84.89 ± 0.11	86.91 ± 0.10	87.53 ± 0.10	88.08 ± 0.10
ProtoNet [13]	63.73 ± 0.21	79.40 ± 0.14	82.83 ± 0.12	84.61 ± 0.11	85.07 ± 0.11	85.57 ± 0.10
+ LASTSHOT (NC)	64.76 ± 0.20	81.60 ± 0.14	85.03 ± 0.12	86.94 ± 0.11	87.56 ± 0.10	88.23 ± 0.10
+ LASTSHOT (LR)	64.85 ± 0.20	81.81 ± 0.14	85.27 ± 0.12	87.19 ± 0.11	87.89 ± 0.10	88.45 ± 0.10
FEAT [23]	66.78 ± 0.20	82.05 ± 0.14	85.15 ± 0.12	87.09 ± 0.11	87.82 ± 0.10	87.83 ± 0.10
+ LASTSHOT (NC)	67.33 ± 0.20	82.39 ± 0.14	85.64 ± 0.12	87.52 ± 0.11	88.26 ± 0.10	88.76 ± 0.10
+ LASTSHOT (LR)	67.35 ± 0.20	82.58 ± 0.14	85.99 ± 0.12	87.80 ± 0.11	88.63 ± 0.10	89.03 ± 0.10



在MiniImageNet数据集尝试每类1-50样本分类任务

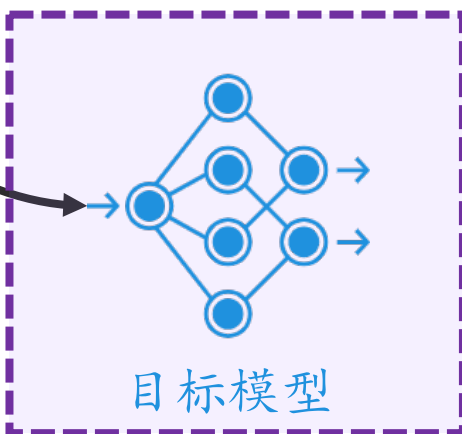
总结和展望



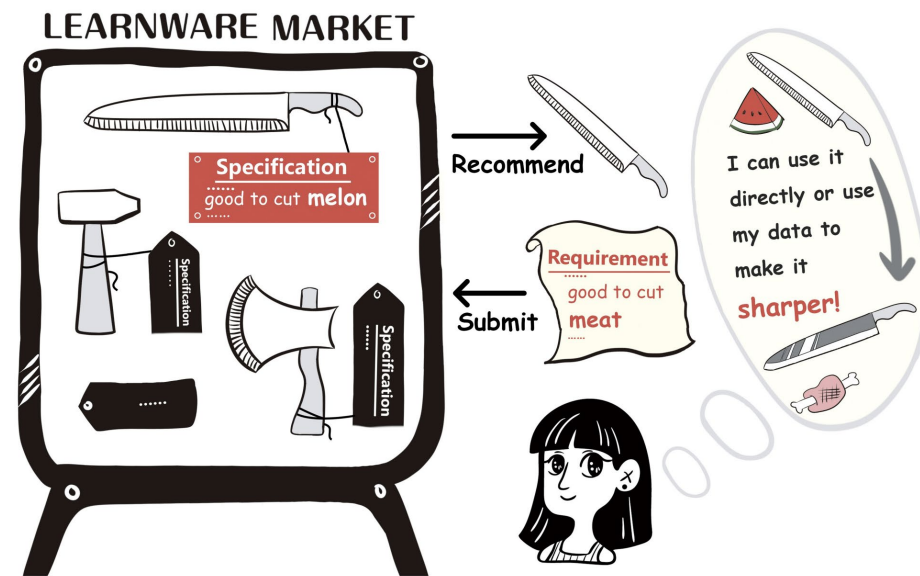
学习可复用的分类器和特征表示，
能够用于不同类型的任务

可复用（Reusable）

解耦方式从相关模型
中抽取有效“经验”



复用（Reuse）



[Zhou, FCS'16][Zhou and Tan, ArXiv'22]

Joint work with Prof. Zhi-Hua Zhou, Prof. Yuan Jiang, Prof. De-Chuan Zhan,
Prof. Wei-Lun Chao, Su Lu, Lu Ming

谢谢