同Hou[1]类似，Wu[11]等人通过在模型最后一层的全连接分类层添加一个偏置学习层来平衡新旧类之间的差距。对于学习过n类旧类以及m类新类的增量阶段S^t。Wu等人保存了旧类（$1,\ldots,n$）类中分类器的特征输出，对新的样本类别使用一阶线性偏置层。即：

$$q_k = \begin{cases} O_k & ,1 \le k \le n \\ \alpha O_k + \beta & ,\ n+1 \le k \le n+m \end{cases}$$

α和β为偏置层的参数，在每个训练阶段完成样本训练之后进行更新，即在每个增量阶段S^t中α和β被所有类共享，对于不同的增量阶段S^g,α和β则不相同。

2.2 问题定义

与传统的类增量学习所假设的类通常是均衡的不同，现实中的大多数数据集合往往是长尾分布的。为了区别于一般的类增量学习环境以针对长尾环境下的类增量学习做进一步研究，我们对长尾环境下的类增量学习做如下规范。

考虑一个用于类增量的流式有标注数据集$D = \{C^1, C^2, C^3, \ldots, C^{n-2}, C^{n-1}, C^n\}$，其中$C^i = \{(x_j^i, y_j^i)\}_{j=0,\ldots,|C^i|}^{|C^i|}$表示一类样本的所有数据。使用$S^t = \{C^{k(t-1)}, \ldots, C^{k*t}\}^t$表示一个增量阶段可见的训练样本，其中$k$表示每个增量阶段所包含的类别数量。不同的增量阶段要求$\forall i,j,\ S^i \cap S^j = \emptyset\ (i \ne j)$。对于不同的$C^i$，其数据量往往不相等，在每个$S^t$中，每个类的数据量分布表现为指数下降的长尾数据模式（如图表1）。同时我们确保在数据类中存在一些类C^i，其数据量$5 \le |C^i| \le 10$的少样本类，用以衡量在极端长尾环境中模型的表现。

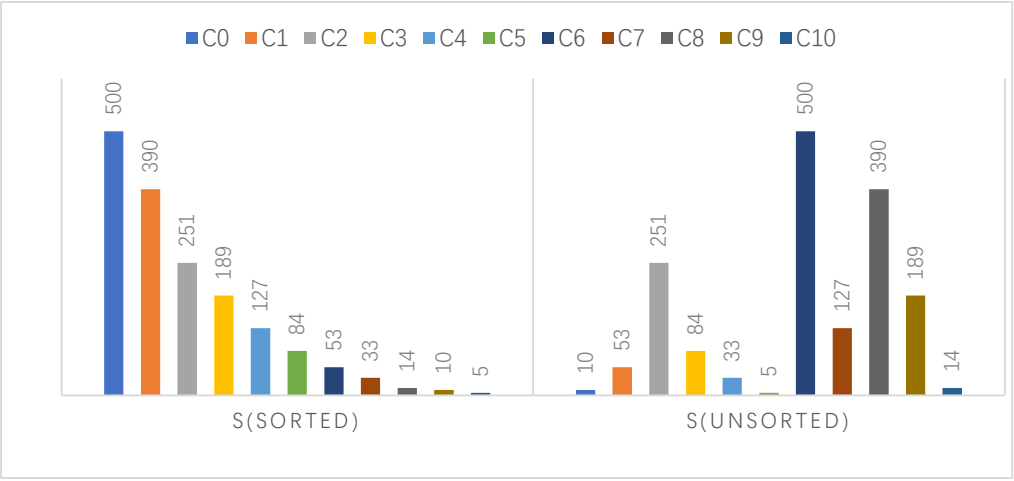

图 2 增量阶段数据分布

假设存在一个符合上述数据规范的数据流，在整体上样本数量符合指数下降函数$s = {}^{1}/_{R^i}$(i为类的序号)的分布规律，在对其进行随机排序后，我们能够认为在所描述的增量阶段S^t中的数据样本数量也整体符合指数下降的分布规律。在一个所有增量阶段为$S = \{S^1, S^2, \dots, S^T\}$的训练中，模型仅能够接触到当前增量阶段$S^t = \{C^{k(t-1)}, \dots, C^{k*t}\}^t$中的所有数据，以及额外内存中的样本数据（如果有）。在每个增量阶段S^t结束后，模型会在包含过往所有类别的测试集$Test = \{T^1, T^2, \dots, T^{k*t-1}, T^{k*t}\}$上进行评估，其中$|T^1| = |T^2| = \cdots = |T^{k*t-1}| = |T^{k*t}|$。我们称其为长尾增量数据集（Long-tail Increamental Dataset,LID）。

在此基础上，我们的目的是建立一个学习器，能够在LID上自动的学习新的分类概念，最终达到良好的分类效果。学习器允许配备一个小容量的额外内存用于存放过往样本（以往的研究过做指出这样能达到更好的效果），学习器可能会拥有多个可调的参数和超参数，在LID上学习的过程中可以自动或手动被更新。

在训练过程中，模型Θ按照增量阶段$1,2,\dots,m-1,m$依次增量的在$S^1, S^2, \dots, S^{m-1}, S^m$上训练，增量阶段I（stage I）模型无法再次范围以前的数据，只能在S^I所包含的数据类上训练。在所有训练阶段完成后，模型M将在包含所有数据类别的测试集$T = \{C_t^1, C_t^2, \dots, C_t^{n-1}, C_t^n\}$上进行测试，在测试集中每个类是均衡的，即保证$|C_t^1| = |C_t^2| = \cdots = |C_t^{n-1}| = |C_t^n|$。在测试阶段所有在测试集$T$中的数据可以被模型多次访问重复测试。

本研究所评价的模型最终表现结果由在最后一个训练阶段S^T上测试平均准确率衡量。为了保证实验的可信程度，可以使用多次验证取平均值的方式展现。同时为了更进一步的展现模型的可用性，我们会在本问题上应用传统的类增量学习模型，并以图示的方式展现对比结果。

2.3 改进方法

为了实现长尾环境下的类增量学习，我们定义了一个4层的CNN网络作为特征提取器$f(\theta)$ 来对数据图像进行特征提取。特征提取器$f(\theta)$定义了特征空间$F \subseteq R^n$，与类增量学习通常使用*softmax*函数来对每一个特征向量P_i进行分类不同，我们使用在特征空间中直接基于度量的分类方式，来避免最后一层分类器在增量阶段参数更新的过程引发灾难性遗忘。

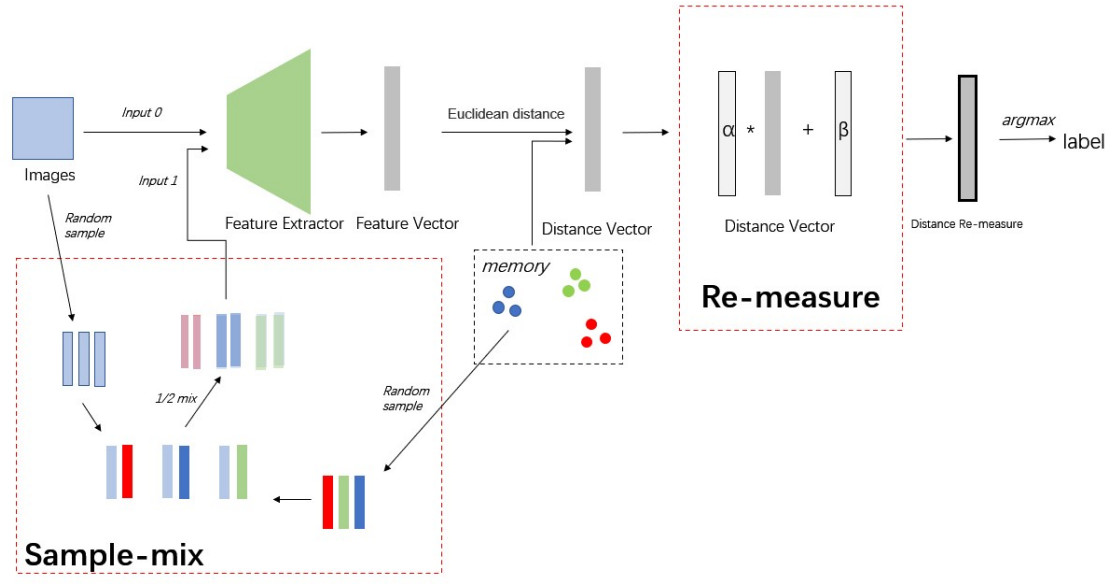

图 3 解决框架中 Sample-mix 与 Re-measure 的图示。memory 中每个类别存储固定数量的样本，通过随机选取样本子集与输入训练图像做 Sample-mix。测试图像完成计算与各类中心的距离之后将距离拼接成一个向量，通过 Re-measure 操作重新度量与各类中心之间的距离之后取最小距离的类的标签作为测试图像的预测值。

每个类的训练图像经过特征提取器之后，会与内存中保存的样本图像计算NME来进行预测。通过设置一个超参数μ表示内存中每类存储样本的数量。为了消除长尾数据带来的类间不均衡的影响，我们提出了使用样本混合(mix)加重新衡量类间距离（Re-measure）的方法来抹平数据间类间不均衡的影响。而对于类增量学习带来的灾难性遗忘的问题，我们尝试通过设计多个扩大类间距离和减少跨增量阶段遗忘的损失函数来避免知识遗忘。为了进一步详细的解释网络模型的细节，我们接下来的所有讨论都基于增量学习中的以下环境：我们假设在增量阶段S^t开始的时候模型已经完成了在n个类别上的训练任务，并即将在$m \subseteq S^t$类的类别样本上训练，此时内存中已经存在n个类别的特定数目的样本。

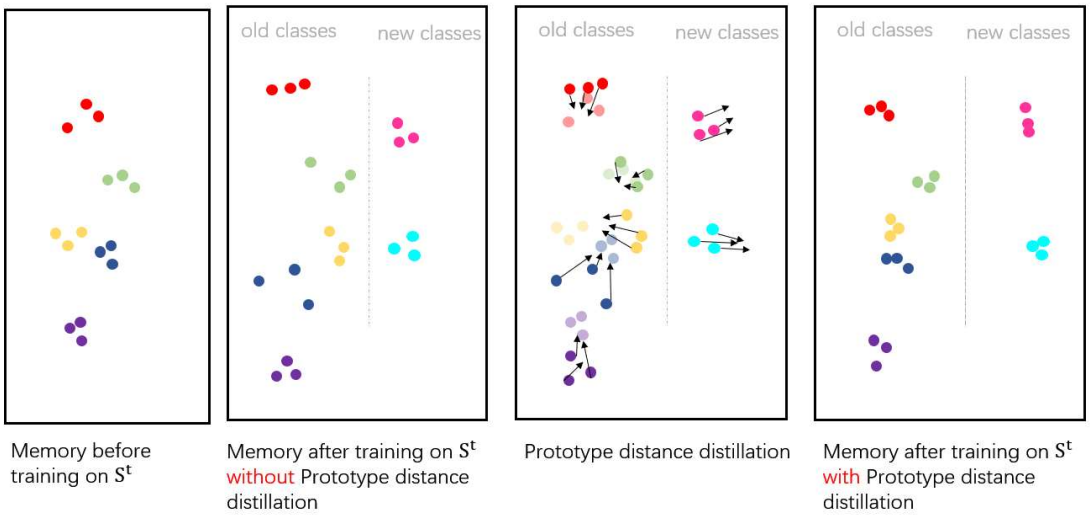

图 4 原型距离蒸馏（Prototype distance distillation）图示

我们接下来对本文绪论中提及的三个改进措施分别做具体描述。

2.3.1 样本混合（Sample mix）

如图3所示，我们的模型设置了一块额外的内存区域用于存放过往的样本，通过设置一个超参数δ来指定内存中每个类别所存储的样本数，即目前共有$\delta * n$个样本。对于新训练的m类样本，我们采用了与iCaRL类似的NME的方式来进行样本选择。对于$C^i \in m$，计算类内中心$C^i_{mean} = \sum_{k=1}^{|C^i|} f(\theta; x^k)/|C^i|$，选取与类内中心最近的$\delta$个样本作为支持样本存入内存。

$$y^* = \underset{y=1,\ldots,|C^i|}{argmin}^{\delta} ||\theta(x) - C^y_{mean}||$$

此时内存中的图像一方面被用于通过特征提取器$f(\theta)$来对每个类别计算类别中心以支持分类。另一方面为了避免长尾数据带来的类间偏置，我们提出可以使用内存中存储的图像与当前训练图像进行随机样本混合，以达到增大类间间距和避免过拟合的效果。

对于新输入的m类样本，模型随机在输入样本和内存样本中各选择一个切片进行样本融合，样本融合采用将两个图像和标签同时平均的方式来增强数据，具体如下：

Algorithm 1 Sample mix

Input $C^{n+1}, C^{n+2}, \cdots, C^{n+m}$ //training class examples
Input $M^1, M^2, \cdots, M^n$ //memory class examples
Require γ //number of slice sample
P ← RandomSlice ($C^{n+1}, C^{n+2}, \cdots, C^{n+m}$; γ) // randomly selected slice containing γ sample
PM ← RandomSlice ($M^1, M^2, \cdots, M^n$; γ)

for i = 1,2,⋯, γ **do**
$x_P^i,\ y_P^i \leftarrow P^i$ //image and label in P
$x_{PM}^i,\ y_{PM}^i \leftarrow PM^i$ //image and label in PM
$x_{Mix}^i \leftarrow (x_{PM}^i + x_P^i)/2$ //average on image
$Mix^i \leftarrow [(x_{Mix}^i, y_P^i), (x_{Mix}^i, y_{PM}^i)]$
end for
output $Mix = \{Mix^1, Mix^2, \ldots, Mix^\gamma\}$

图 5 **样本融合算法**

最后将融合后的数据集Mix作为训练集的一部分通过模型，以此来更新网络参数。由于输入数据是不均衡（long-tail）的，而内存中存储的样本是经过人工均衡的，我们认为样本融合在一定程度上为输入数据提供了均衡程度，并帮助平衡了历史学习数据与当前学习样本之间的偏差。

2.3.2 重新度量（Re-measure）

我们认真回顾了Prototypical Networks[15]的部分工作，考虑采用在特征空间中基于欧式距离度量的方式为测试样本进行分类。但我们的实验结果表明，如图6所示，在一个独立的增量阶段S^t中，模型对于不同类别C^i和C^j的分类效果存在明显偏差，其准确率与类样本数量存在较强的相关性，这种偏差很有可能是长尾数据导致的原生类间不均衡。

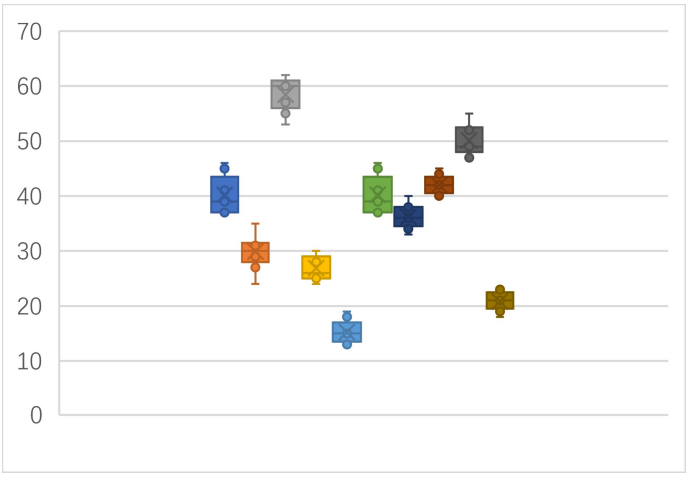

图 6 增量阶段中不同类别的分类准确率（10 次测试平均值）

针对这种增量阶段内的类间偏差，我们提出基于重新度量（Re-measure）的方法，来在一定程度上抹平这种增量阶段内的类间不均衡。在模型中我们添加一一个重新度量层，由于长尾数据中包含大量的“尾部”类，为避免发生重新度量层的过拟合，我们使用线性模型来学习这种偏差。模型通过每个类别分配两个可学习的参数α和β。在C^i计算完特征向量与每个类的中心距离$dist = \{d_i^1, d_i^2, \dots, d_i^{n+m}\}$后，重新度量类间距离：

$$dist_{re}^i = \alpha d_i^j + \beta \quad (j \in 1, \dots, n+m)$$

其中d_i^j表示样本i与j类的类别中心之间的欧式距离。再计算完$dist_{re}^i$之后再由NCM算法计算出样本i的预测标签。

由于重新度量层在不同的训练阶段使用不同的参数（由于为每个类别单独分配了不同的α和β），我们在每个训练阶段单独训练当前重新度量层的参数。在后续增量阶段，前类别的度量层参数被冻结以保证增量阶段间的相互独立性。

2.3.3 原型距离蒸馏（Prototype Distance Distillation）

传统的类增量学习面临着灾难性遗忘的困扰，长尾环境下的类增量学习也是如此。我们前面所介绍的具有代表性的类增量学习方法中，大部分都应用了样本重现加知识蒸馏损失函数的方式。由于传统类增量学习模型使用全连接层加softmax函数计算概率输出，而我们的模型使用基于NCM的距离分类，传统的类增量学习方法并不能够直接适用于本实验场景。在实验过程我们意识到更紧凑的类内距离以及更广阔的类间距离能够帮助模型提高分类结果，为了避免灾难性遗忘的同时保证模型在新类上的学习效果，我们应该在模型训练过程中尽

可能地紧凑类内距离，扩大类间间距，其中扩大类间间距2.2.3节的样本混合方法起到了一定的效果，由于样本混合在一定程度上增加了数据量，在此基础上我们通过损失函数控制类间间距可以较少的考虑过拟合的影响。

我们基于本模型以NCM来分类的方法设计了原型距离蒸馏（Prototype distance distillation）。计算方法基于如下情景，在S^{t-1}训练完成后，我们从内存中提取样本$\{C_{mem}^1, C_{mem}^2, \dots, C_{mem}^n\}$计算n类的类内中心$\{O^1, O^2, \dots, O^{n-1}, O^n\}$，在增量阶段$S^t$的训练过程中，我们再根据内存样本$\{C_{mem}^1, C_{mem}^2, \dots, C_{mem}^n\}$通过特征提取器计算出新的(n)类的类内中心$\{N^1, N^2, \dots, N^{n-1}, N^n\}$，当前输入的（m）类数据通过特征提取器计算出类内中心$\{N^{n+1}, \dots, N^{n+m-1}, N^{n+m}\}$，为了衡量和减少两个训练阶段之间的差距同时为了更好地在新m类上分类，我们按照如图7所示的算法分别计算引力损失（Pull Loss）和斥力损失（Push Loss）：

Algorithm 2 Prototype Distance Distillation

Input $O^1, O^2, \dots, O^{n-1}, O^n$ //old class center prototype
Input $N^1, N^2, \dots, N^n$ //new class center prototype compute from memory
Input $N^{n+1}, \dots, N^{n+m-1}, N^{n+m}$ //new class center prototype compute from input
Require T //hyper-parameter

$Dist_{pull} \leftarrow \{ \| O^1 - N^1 \|, \| O^2 - N^2 \|, \cdots, \| O^n - N^n \| \}$
$Dist_{push} \leftarrow \{ \| O^i - N^j \| \} (\forall i \neq j, i \in (1,n), j \in (1,n+m))$
$Dist_{pull} \leftarrow Dist_{pull} / \mathrm{T}$
$Dist_{push} \leftarrow Dist_{push} / \mathrm{T}$
$Pull\ Loss \leftarrow softmax(Dist_{pull})$
$Push\ Loss \leftarrow 1/\sqrt{Dist_{push}}$
output *Pull Loss, Push Loss*

图 7 Prototype Distance Distillation **计算方法**

损失函数的另一个组成部分是训练过程中分类产生的误差损失$Loss_{Error}$，我们分别实现了基于NME的损失函数以及基于交叉熵的损失函数以比对那种误差损失的形式更适用于本实验环境。

$$Loss_{NME} = softmax\left(-\left\{\left|\left|n_i - N^j\right|\right|\right\}\right) - \log\_softmax\left(-\left\{\left|\left|n_i - N^i\right|\right|\right\}\right)$$

$Loss_{CrossEntropy}$

$$= softmax\left(-\left\{\left||n_i - N^j|\right|\right\}\right) - \sum_i y^i \, log\hat{y}^i + (1 - y^i)\log(1 - \hat{y}^i)$$

$$\forall i, j \in (n+1, n+m), i \neq j$$

其中y^i代表i类样本的标签。

3 实验

我们基于 2.2 所述的问题规范进行对照试验，并与 2.1 节所述的类增量学习基线模型进行对比。我们基于 Cifar-100 数据集[10]实现了用于本实验测试的数据集。数据集的详细情况如下：

Cifar-100数据集：原始数据集包含60,000张32x32的三通道RGB图像，包含100个类，每个类包含500张训练图像和100张测试图像。我们将100个类别分为5个增量阶段，每个增量阶段包含20个类别。对于测试图像我们不做调整，即在测试阶段每个类别都拥有100张测试图像。我们对训练图像集中的样本使用指数下降函数$f = 1/R^i$进行随机抽取，其中R为可变参数，i为随机数。最终抽取出的train数据集在整体上符合指数分布。为了测试多种情况，我们分别使用了函数$1/20^r$，$1/30^r$和$1/40^r$（r代表随机数）进行了样本筛选。

增量阶段样本数	**1**	**2**	**3**	**4**	**5**
$\mathbf{1/20^r}$	3227(个)	3196	3145	3143	3162
$\mathbf{1/30^r}$	2903(个)	2833	2821	2832	2906
$\mathbf{1/40^r}$	2647(个)	2642	2639	2601	2654

表格 1 不同筛选函数下每个增量阶段的样本数

实验设置方面，我们将训练集/测试集的分割比例设置为9：1，由于在训练过程中内存存储和样本融合机制的存在，实际训练样本会更多，并要求更大的训练内存。

3.1 实现细节

我们的模型使用PyTorch实现，分别使用了ResNet32[18]以及简单的4层卷积神经网络作为特征提取器用以对比。模型的特征提取器在每个训练阶段训练60个epochs，学习率被初始设置为0.001，并在每训练20次（即20个epochs）之后变为原来的0.5倍以保证增量过程中更少的发生灾难性遗忘和过拟合。重新度量层在每个训练阶段在与特征提取器相同的训练数据上训练3个epochs，学习率被初始化设置为0.0001，并在每10个epoch之后变为原来的2倍，以保证在学习到增量阶段内的数据不平衡的同时也能够学习到增量阶段之间的不平衡。在数据预处理阶段我们沿用了ResNet[18]中的处理方式使用了随机裁剪和随机水平翻转来做数据增强。由于长尾数据分布的影响，用于控制内存容量的超参数μ不应该大于整个数据集中最少类的样本数量，以防止支持集不足导致模型对于小样本类分类的系统误差。本实验中δ=5，即每各类别在内存中被保存5个样本。

我们将用此方法与最先进的类增量学习模型NCM[1]和BiC[11]对比在长尾增量数据集上的表现。由于BiC和NCM模型在设计上都设计了额外内存机制，为了与本实验提出的框架相统一，我们将所有被测模型的内存的最大容量都设置为了δ*总类别数=5*100=500。

3.2 对照实验

我们分别在符合概率分布$1/20^r$，$1/30^r$和$1/40^r$（r代表随机数）的长尾数据集上进行了类增量学习的测试，结果如图7所示。

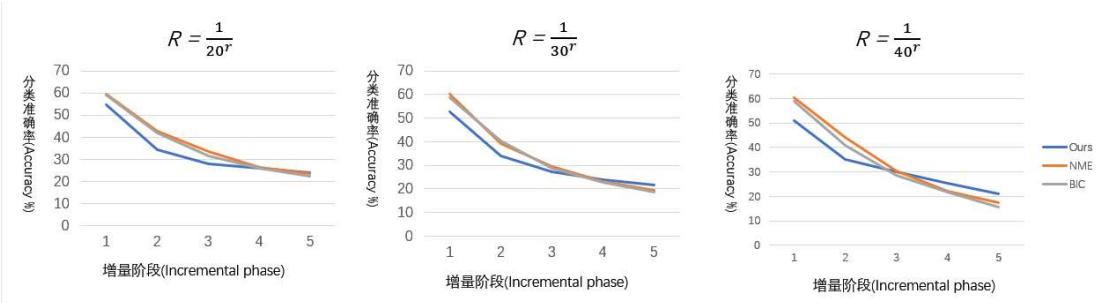

图 8 BiC，NME 和我们的模型在不同的长尾分布上的对照实验结果。R 表示数据集样本数量的指数分布，其中底数越大代表整体数据分布越不均匀，总样本数量越少。

在长尾数据集三种不同分布下，我们的模型在最终的测试结果（图中增量阶段 5）上，优于已有最先进的类增量学习算法。在数据分布$R = 1/20^r$中，我们模型的准确率比 NME 和 BiC 算法的准确率分别高 0.6%和 0.9%。在数据分布$R =$

$1/30^r$中，我们模型的准确率分别比 NME 和 BiC 算法高 2.3%和 3.2%。在数据分布$R = 1/40^r$中，我们的模型分别比 NME 和 BiC 算法高 3.6%和 5.4%。

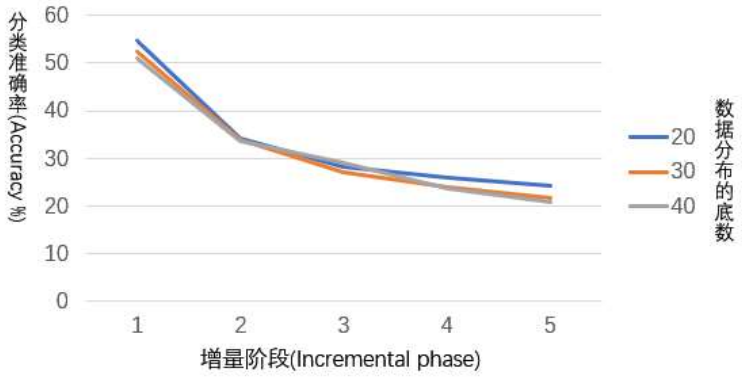

图 9 不同数据分布对我们模型准确率的影响

如图 8 和图 9 所示，我们通过实验验证了，数据集合的长尾分布对模型的分类准确率存在较大影响。同时实验结果体现出了，数据表现出越极端的长尾分布（数据分布函数的底数越大），对模型的影响性就越强。

注意到图 7 中，长尾数据分布越极端，对传统类增量学习方法的影响和对本研究提出的方法的影响差距越大。由于传统类增量方法并没有考虑如何处理原生增量数据中存在的不均衡问题，出现这样的结果在长尾分布数据正常的影响范围之内。注意到类增量学习开始的时候，传统类增量学习模型大幅由于本研究提出的算法，由于采用了相同的特征提取器，且在第 1、2 增量阶段不能存在明显的灾难性遗忘，我们认为出现这样的情况是由于 NCM 分类方法和传统全连接层加 softmax 计算概率输出之间的差异。由于基于欧氏距离在特征空间进行分类的局限性，在数据充足的情况下，仅通过距离来衡量分类未必能学习到充分的特征以及拥有很好的泛化性。这是由于基于欧式距离的度量$dist = \left|\left|C^i - C^i_{mean}\right|\right|$，这代表只有该类的所有样本都被特征提取器映射到一个以C^i_{mean}为中心的高维单位球体才能够正确分类。由于高维特征空间的分布特性，基于单位球的分类方法在样本数量密集的时候并不能够很好的分割特征空间，从而导致样本的分类误差。由此在开始的增量阶段，NCM 的分类方法准确率要大幅低于基于全连接层加 softmax 函数分类的方法。

Method	Incremental Phase（增量阶段）R= $\frac{1}{40^r}$				
	1	2	3	4	5
Joint CNN	60.75 (%)	54.75	51.5	48.7	44.8
BiC	58.85 (%)	40.75	28.65	21.83	15.87
NME	60.41 (%)	44.12	30.31	22.08	17.35
Ours	51.08 (%)	35.01	29.97	25.32	20.91

图 10 **在** R= 1/40^r **分布的情况下各方法准确率对比**

随着增量阶段的增长，传统类增量学习方法受到的影响远大于本研究提出的方法所受到的影响（如图 10）。我们认为这是由于长尾分布数据进一步的“开放”给训练模型，导致数据类样本类间的不均衡被放大，而本研究提出的方法提供了重新距离度量（Re-measure）和样本融合的方法在一定程度上均衡了类间样本数量上的差距，同时非参数化的 NCM 分类方法也比传统的全连接方法提供了更好的抗遗忘性。

Loss Function	Incremental Phase（增量阶段）R= $\frac{1}{40^r}$				
	1	2	3	4	5
NCM	51.08 (%)	35.01	29.97	25.32	20.91
Cross Entropy	44.38 (%)	34.48	28.85	20.25	15.57

图 11 **在数据分布** R=1/40^r **的情况下，**NCM **损失函数与** CrossEntropy **损失函数的对比结果**

我们在数据分布符合 $R=1/40^r$ 的情况下测试了最后分类误差损失函数 $Loss_{NCM}$ 和 $Loss_{CrossEntropy}$ 的差距，结果如图 11 所示。在本实验中使用 *NCM* 计算分类误差损失函数的效果要远好于使用 *CrossEntropy* 计算。

4 切除分析

长尾环境下的类增量学习与传统类增量学习的定义相比，引入了原始数据分布不均衡的问题，在此基础上，样本的数量会对模型最终的分类结果产生较大的影响。我们已经分析过，传统模型在长尾环境下容易产生“歧视性偏差”

以及针对小样本老类产生过拟合的问题。为了在新的问题环境下解决这些问题，我们分别提出了：(1)使用NCM距离度量代替传统全连接加softmax形式的分类器。(2)引入了样本融合机制。(3)引入了距离重新度量机制。(4)改进基于概率输出的蒸馏损失为基于特征中心绝对位置的距离损失。接下来我们将结合理论考量和实验结果分析所应用的方案的有效性。

4.1 距离重新度量切除

基于平衡样本增量阶段内偏差以及增量阶段间数据偏差的想法，我们提出使用一阶线性函数对分类距离进行学习来平衡由于样本数量导致的偏置。

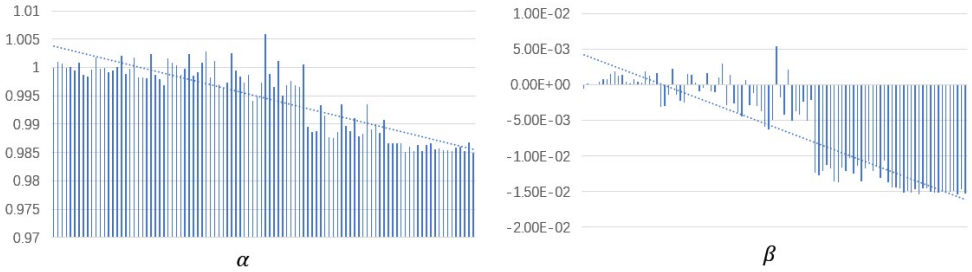

图 12 距离重新度量层的参数，增量阶段=5，数据分布 R=$1/40^r$

如图 12 所示是在数据分布 R=$1/40^r$的分布下，距离度量层经过训练之后的参数。我们可以明显的观察到α和β参数存在明显的随类别 id 下降的趋势。这表现出距离重新度量层对较后的增量阶段样本存在较大的距离缩小纠正，即距离重新度量层学习出了模型分类对新类的偏置，以平衡新类和老类的距离差异。这与以往的研究结果[1,11]相符。同时在一个独立的增量阶段内部，我们观察到α参数的大小与样本数量呈现负相关的关系。

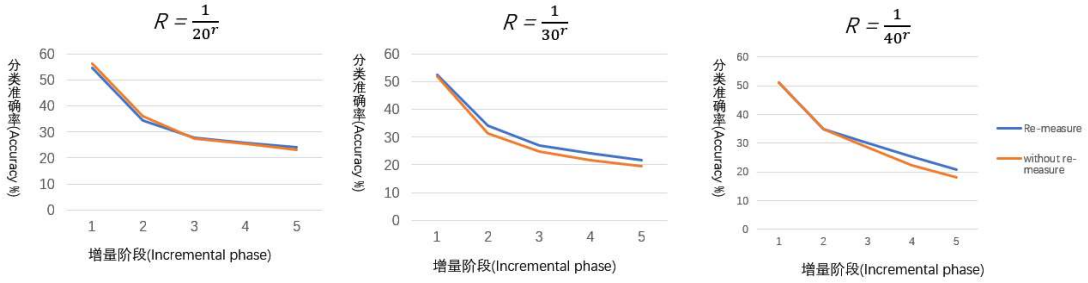

图 13 在不同的数据分布下，距离重新度量层的切除实验结果

在冻结距离重新度量层的层数更新（即保持α和β为 1 和 0 以消除重新度量层的影响）之后，我们观测到模型最终的准确率大幅下降，并且观察最后增量阶段的分类准确率发现了较大的类间差距。如图 13 所示在长尾分布分别符合$R=1/20^r$，$R=1/30^r$，$R=1/40^r$的情况下，重新度量层分别提升了模型最终分类准确率 1%，2.1%和 2.9%。在分析实验数据的基础上我们认为，距离重新度量层帮助平衡了增量阶段内的类间差异以及增量阶段间的类间差异。对于长尾环境下的类增量学习来说，重新度量距离是必要的。

4.2 样本融合切除

样本融合被设计用来解决长尾分布数据带来的原生的数据分布不均衡。通过内存样本和输入样本的随机选取，进行样本融合，我们提升了模型在最后增量阶段的分类准确率，结果如图 14 所示。

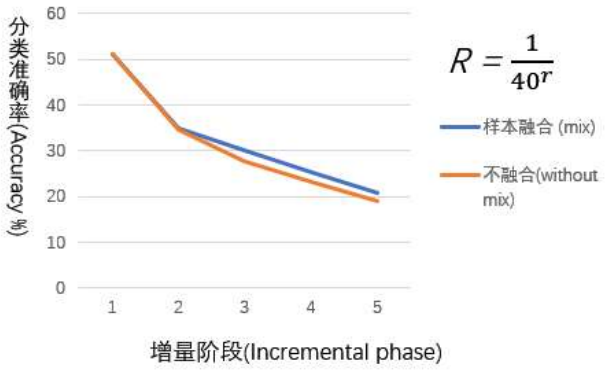

图 14 在数据分布 R=1/40^r 的情况下，样本融合的切除实验结果

在长尾数据分布遵循 $R=1/40^r$ 的情况下，样本融合的方法在最终增量阶段 5 上的准确率相比不使用样本融合提高了 1.3%。

4.3 原型距离蒸馏切除

原型距离蒸馏（Prototype Distance Distillation）被设计用来减少跨增量阶段的知识遗忘，同时扩大类间特征中心距离。在实验中这两项损失被设计为 Push Loss（用于扩大类间特征中心距离）和 Pull Loss（用于距离蒸馏）。

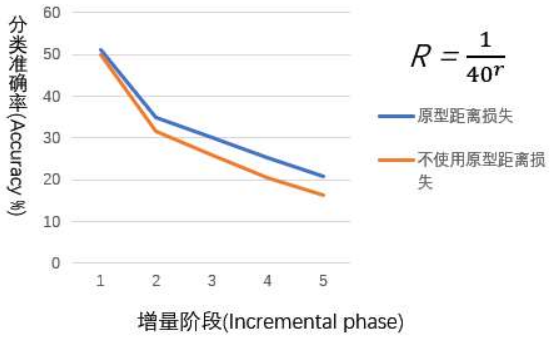

图 15 在数据分布 R=1/40^r 的情况下，原型距离损失的切除实验结果

在长尾数据分布遵循 $R=1/40^r$ 的情况下，原型距离蒸馏的方法对模型准确率的影响随增量阶段而增加，在增量阶段从 2 到 5，准确率分别提升了 3.5%，4.2%，4.8%，4.7%。

5 结论

虽然深度学习在过去取得了巨大的成就并成功的融入人们的生活，但是其训练过程要求的大量的、均衡的数据在现实中不能够被快速大量的提供。在更多情况下，数据是长尾分布且增量获得的，这为传统深度学习模型在现实中的应用和部署提出了包括灾难性遗忘等在内的很多挑战。传统类增量学习的方法对应对这些挑战有一定的帮助，但在更极端的长尾环境下这些方法的效果也会大幅下降。

本文在研究这些挑战的情况下给出了长尾环境下的类增量学习的问题定义。通过改造常用数据集构造了长尾的“Cifar-100”测试基准，并在此基准上提出了三种基于特征距离分类的改进方法。我们的实验表明通过重新衡量距离、样本融合以及针对特征原型的知识蒸馏的改进，可以提供超出传统深度学习模型的巨大准确率提升。我们的研究只是解决深度学习模型更广泛应用等问题的其中一部分，我们认为本论文的研究与尝试能够为之后的探索提供实践基础以及思路指引。

参考文献

[1] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 831–839, 2019.

[2] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2001–2010, 2017.

[3] Davis Wertheimer, Bharath Hariharan. Few-Shot Learning with Localization in Realistic Settings. arXiv preprint arXiv: 1904.08502, 2019

[4] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. ´ Zitnick. Microsoft COCO: common objects in context. In European Conference on Computer Vision, 2014.7

[5] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[7] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Neural Information Processing Systems, pages 487–495. Curran Associates, Inc., 2014.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[10] A. Krizhevsky. Learning multiple layers of features from tiny images. Master’s thesis, Department of Computer Science, University of Toronto, 2009.

[11] Wu Yue, Chen Yinpeng, Wang Lijuan, Ye Yuancheng, Liu Zicheng, Guo Yandong, and Fu Yun. Large scale incremental learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[12] Yuxiong Wang, Deva Kannan Ramanan and Martial Hebert. Learning to Model the Tail. International Conference on Neural Information Processing Systems (NIPS '17), pp. 7032 – 7042, December, 2017

[13] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, Wenhui Li. Deep Representation Learning on Long-tail Data: A Learnable Embedding Augmentation Perspective. arXiv preprint arXiv:2002.10826, 2020.

[14] Dvir Samuel, Yuval Atzmon, Gal Chechik. Long-tail learning with attributes. arXiv preprint arXiv:2004.02235, 2020.

[15] Zhizhong Li, Derek Hoiem. Learning without Forgetting. arXiv preprint arXiv:1606.09282

[16] Jake Snell, Kevin Swersky, Richard S. Zemel. Prototypical Networks for Few-shot Learning.

NIPS2017. arXiv preprint arXiv: 1703.05175,2017.

[17] Lu Yu1,2 , Bartomiej Twardowski2 , Xialei Liu2 , Luis Herranz2 , Kai Wang2 , Yongmei Cheng1 , Shangling Jui3 , Joost van de Weijer2. Semantic Drift Compensation for Class-Incremental Learning. arXiv preprint arXiv: 2004.00440, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.

[20] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016.

[21] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In Advances in Neural Inf

[22] Amal Rannen Ep Triki, Rahaf Aljundi, Matthew Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In Proceedings ICCV 2017, pages 1320–1328, 2017.

[23] Xing Wei, Yue Zhang, Yihong Gong, Jiawei Zhang, and Nanning Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In The European Conference on Computer Vision (ECCV), September 2018.

[24] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In Advances In Neural Information Processing Systems, pages 5962–5972, 2018.

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 212–220, 2017.

[26] Xialei Liu, Marc Masana, Luis Herranz, Van De Weijer Joost, Antonio M. Lopez, and Andrew D. Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. arxiv preprint arXiv:1802.02950, 2018.

[27] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In Proceedings of the IEEE International Conference on Computer Vision, pages 2759– 2768, 2019.

[28] David Lopez-Paz et al. Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems, pages 6467–6476, 2017.

[29] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Proceedings of the 34th International Conference on Machine LearningVolume 70, pages 3987–3995. JMLR. org, 2017.

[30] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. arXiv preprint arXiv:1708.01547, 2017.

作者简历

姓名：徐亮　性别：男　民族：汉　出生年月：1998-01-05　籍贯：江苏省镇江市

2013.09-2016.06　甘肃省嘉峪关市酒钢三中

2016.09-2020.07　浙江大学攻读学士学位

获奖情况：第三届 BitRun 区块链黑客松大赛二等奖，浙江大学程序设计竞赛三等奖。

参加项目：腾讯课堂后台监控系统搭建，区块链钱包 Super Wallet 开发

本科生毕业论文（设计）任务书

一、题目：长尾环境下的类增量学习

二、指导教师对毕业论文（设计）的进度安排及任务要求：

本论文的目标旨在提出新的增量学习算法，以对增量学习的理论发展和实际应用做出一定的贡献。在文献综述整理增量学习领域的研究现状的基础上，提出长尾环境下类增量学习的问题定义。理解长尾环境下类增量学习的研究背景，目的及意义。明确研究目标和具体的研究内容。构造长尾环境下类增量学习的相关数据集，并使用基准算法进行比较实验。分析总结出长尾环境所带来的困难和挑战。提出可行的技术路线。在实现路线上，首先做一定实验确定问题是否有研究的意义，并在此基础上通过适度的理论创新，尝试提出一种解决方案。如果解决方案与已有研究相比取得了一定的优势，再通过对比实验和切除分析来讨论，确定论文的研究成果。

起讫日期 2019 年 9 月 15 日 至 2020 年 6 月 15 日

指导教师（签名）_____职称_____

三、系或研究所审核意见:

负责人（签名）_____

年 月 日

毕 业 论 文（设计） 考 核

一、指导教师对毕业论文（设计）的评语：

论文针对长尾分布的类增量学习问题，提出基于特征距离的分类模型的三点改进： 使用样本混合来避免过拟合以及增加类间差距；使用距离重新度量改进对新类的熟悉性偏差及对少类的歧视性偏差；使用原型距离蒸馏使模型减少灾难性遗忘。实验证实了改进方法的有效性。论文逻辑清晰，结构合理，实验详实，是一篇较好的本科毕业论文。

指导教师(签名）

年 月 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述/中期报告占（10%）	开题报告占（15%）	外文翻译占（5%）	毕业论文(设计）质量及答辩占（70%）	总评成绩
分值					

答辩小组负责人（签名）

年 月 日

浙江大学

本 科 生 毕 业 论 文

文献综述和开题报告

学生姓名　　徐 亮

学生学号　　3160105865

指导教师　　孙建伶

年级与专业　　16 级 计算机科学与技术

所在学院　　计算机科学与技术学院

一、题目：长尾环境下的类增量学习

二、指导教师对文献综述和开题报告的具体要求：

1.文献综述要求：

综述近期关于深度学习的类增量学习领域的研究论文，不少于20篇。梳理研究现状，对已有方法进行分类。指出现有方法的不足，并讨论潜在的研究方向。

2.开题报告要求：

理解长尾环境下类增量学习的研究背景，目的及意义。明确研究目标和具体的研究内容。构造长尾环境下类增量学习的相关数据集，并使用基准算法进行比较实验。 分析总结出长尾环境所带来的困难和挑战。提出可行的技术路线，制定合理的研究进度计划，明确最终成果形式（包括完成研究论文）。

3.外文翻译要求：

翻译下列论文，要求译文易读，完整，准确。

《Few-Shot Learning with Localization in Realistic Settings》

指导教师（签名）.....

年　　月　　日

目　录

一、文献综述

1. 背景介绍

近几年的人工智能飞速发展,绝大部分是由于深度学习的研究取得了巨大的成功。在很多领域获得了标志性的成果。1980年，Kunihiko Fukushima 提出了第一个卷积神经网络以来深度学习不断发展。2015年，何凯明和他在微软的团队报告说，他们的模型在对来自ImageNet的图像进行分类时表现优于人类。尽管人工智能已经依靠深度学习在十亿级别的图像识别中击败了人类,然而这些成功是在静态的模型上取得的，这意味着他们需要提前收集好大量的数据、做好数据类的人工平衡,才能使深度学习获得一个较好的效果。然而现实中的大多数情况是，数据不会提前全部准备好，往往是随着时间增量累积的。这种情况让提前收集好大量的数据,然后再进行模型的训练非常困难。同时由于现实应用的实时性要求，模型需要实时学习数据，每次新数据可用时，都需要重新从头启动训练过程。这无疑是对资源的大量浪费。

相比之下，动物和人类的自然认知系统则完全不需要这么做。人类的一生都在不断学习新的知识，虽然可能会逐渐忘记一些旧的信息，但不会完全丧失以前学习过的技能，在这一过程中，人类可能会重新回顾旧的概念，但是这对于保留旧的概念并不重要。

如果没有特殊的措施,使用梯度下降的传统机器学习模型就不能够以这种连续的、增量的方式进行学习：当它们学习现有的新的概念时，有可能会完全的遗忘原有的旧的概念。当研究人员使用一个简单的神经网络被用来模拟连续出现的事件时[23]，就观察到了明显的灾难性遗忘（即在学习后面的概念时，完全忘记了旧概念的知识）。这种现象会导致性能的突然下降，或者在最坏的情况下，旧的知识完全被新的知识所覆盖。为了应对这些问题，传统的机器学习方法使用更大的Batch将数据随机混合来进行训练（Guo et al. (2016), LeCun et al. (2015)），然而这种方法不仅要求更大的计算资源，同时假设了，在训练阶段，所有的样本都是机器可以观察的，而这并不符合实际情况，我们认为现实中大部分的环境数据（如动物分类图片、用户购物信息或语音信息等）并不能满足实时

全部获取的条件。而对于传统模型来说，如果不能够一次性获取全部的数据，就只能够在部分的数据集上进行部分训练，或是每当新数据到来时，对新数据进行序列式的训练，而这分别对应了获取信息不全和旧信息灾难性遗忘的问题。为了解决这类问题，研究人员提出了增量学习的研究方向。

本研究的主题是长尾环境下的增量学习，由于相关文献较少，为保证文献综述部分的重点，在这部分我们主要调研和考虑分类问题上的类增量学习的研究成果。关于本研究的主题，具体的环境引入和问题分析将在开题报告部分中加以说明。

2. 国内外研究现状

2.1 研究方向及进展

增量学习（Incremental Learning）研究的是从数据流中学习的问题，目标是逐渐扩展所获得的知识，并将其用于未来的学习。这些数据流可能来自于不断变化的输入数据(例如，不断变化的图像条件)，或者基于任务的、不断变化的任务(例如，不同的分类问题)。定义增量学习的标准是学习过程的顺序性，即数据是存在时间顺序的，其中只有小部分输入的数据能够立刻获得。这其中研究主要面对的挑战是如何在不发生灾难性遗忘的情况下学习模型，也就是说:随着新任务或领域的增加，以前学习过的任务或领域的性能不应该显著随着时间的推移而下降。广义的增量学习模型期望能够妥善的处理好多种数据流的变化，具体来讲，增量学习在发展过程中演化出了三类，分别是样本增量学习、类别增量学习和特征增量学习。我们在此对各种类型的增量学习进行简要的介绍。

1. 样本增量学习

由于新数据的各种原因，样本的特征值可能会改变，每个类别所占的比例也会不断变化。这些都会影响最后分类的准确率。因此，模型需要确保在现有知识不受巨大变化影响的情况下，通过新样本的增量学习来提取新知识，融合新旧知识以提高分类的准确性。

2. 类别增量学习

类别增量学习面临数据新增类别在之前没有遇见过的问题，在这种情

境下模型最后的输出层的参数都会发生变化，类别增量学习主要目的是识别新类，并将其加入现有的类别集合中，提升大分类的智能和准确性。

3. 特征增量学习

随着数据分布随时间变化，一些数据的特征可能随之显现出来，例如新增的数据显示出来某一个目标类往往随着特定的如红色或者蓝色的背景出现，这些新的属性特征能够将分类提升一个很大的程度，并提升分类准确率。特征增量学习的任务是在现有特征空间的基础上，加入新的属性特征，构建新的特征空间，提升分类准确率。

考虑到聚焦重点的情况下，我们仅关注与本研究课题相关的类别增量学习。

目前实验已经证明，一些朴素的分类算法能够很自然的支持增量学习如朴素贝叶斯，支持向量机，决策树，随机森林，人工神经网络，K-最近邻等。现有的学术进展大部分以这些分类算法为基础，做适量的改造和融合。例如针对文本领域中的增量学习方法，早在 08 年罗等人[2]提出了一种新的加权朴素贝叶斯增量学习方法。它的思想有三个主要的改进。首先设置类置信阈值，严格过滤增量样本；其次，根据分类能力，对系数进行手动动态调整，提高分类精度。最后，利用词频权重公式可以反映出该特征的重要性，克服了不能突出不同特征权重的缺陷。

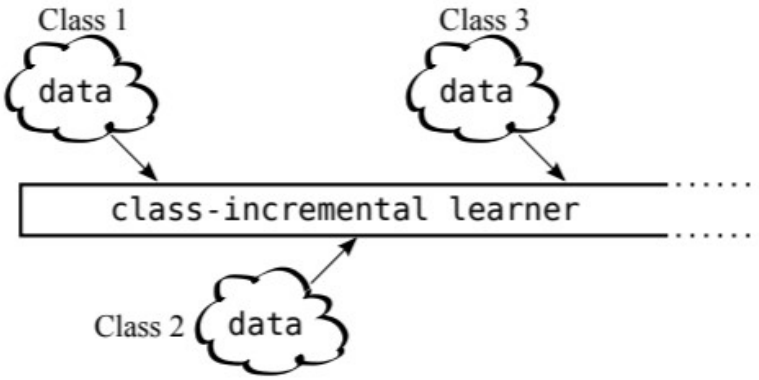

图 1 类增量学习的过程（图源[8]）

现今国内外主流的研究增量学习的方法集中在基于正则的方法和重放

历史样本（Replay-based）上。

基于正则的方法相比重放历史的方法更早地提出，同时相比之下也更容易实现。它们的原理是不存储原始输入，一个重要的工作方向是在损失函数中提出一个额外的正则化项，从而在学习新的知识的时候能够同时巩固原有知识。Learning without Forgetting（LwF）[22]引入了以数据为中心的方法，构造从以前的模型到新数据训练出的模型的知识提炼，即“知识蒸馏方法”，用以减轻遗忘和转移知识，它使用了以前模型的输出作为以前任务的软标签。其他的研究成果[25]也已经介绍了这种方法，然而它已经表明，这种方法只适用于增量差距较小的问题。

基于重放历史样本的方法（Replay-based）在图像分类问题上使用原始格式存储样本或是使用生成模型（GAN）来达到仅储存少量样本的情况下较好的学习效果。这些算法在学习新任务以减轻遗忘的同时，会对模型重放先前任务中存储的样本。这些重放样本可以用于训练，这种方式近似于先前数据和当前数据的混合训练，以限制在学习新类时的梯度下降，从而不干扰以往的学习效果。Rebuffi 等人[8]提出了 iCaRL 的算法，在实验中存储了每个类的样本子集，这个子集被精心选取和淘汰从而子集中心接近被学习的特征空间中每个类的平均值。因为内存是有限的，为了不断适应到来的新类，iCaRL 模型的选择算法会根据距离子集中心的距离来淘汰旧类样本。

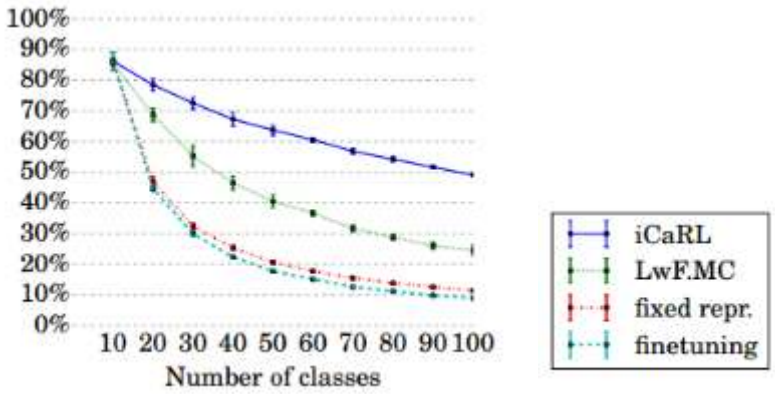

图 2 增量学习 iCaRL 算法相对于传统模型的表现（其中纵坐标代表整体分类准确率，横坐标代表类增量的类数量，蓝线代表的是 iCaRL[8]算法，绿色代表 LwF[22]算法,红色和蓝色是传统模型算法）

由于基于重放的方法可能会导致由于样本增量以及中心不合适选取导

致的过拟合，并且似乎由于重放时间和内存大小收到了一定的限制。使得约束梯度优化成为了一种备选的解决方案。如 GEM[24]中建议的，在当前数据更新的时候，关键仅在于约束当前的梯度下降，从而不干扰旧类别的识别精度。它们通过将当前梯度更新投影到由旧梯度更新所概括出来的可行域上实现的。这保证了在训练过程中不会出现“梯度冲突”，从而保证了训练质量。

大量的研究表明，基于重放的方法相比单纯基于正则的方法拥有更好的识别精度。在近期的研究中，大量的研究人员提出了基于原有类中增量学习模型的很多小技巧[1,7]。Wu[1]等人通过实验发现，在类增量学习中，模型的参数会随着类别数量的增加，输出层的参数会朝向新类偏置，由此导致模型认为一张图片是新来的类的概率非常高。他们提出了 BIC 算法，提出了一种方法，这种算法在输出层前添加一层偏置层，抵消输出层的偏置，克服了新的类和旧的类之间的偏差问题，大幅提升了模型的表现。

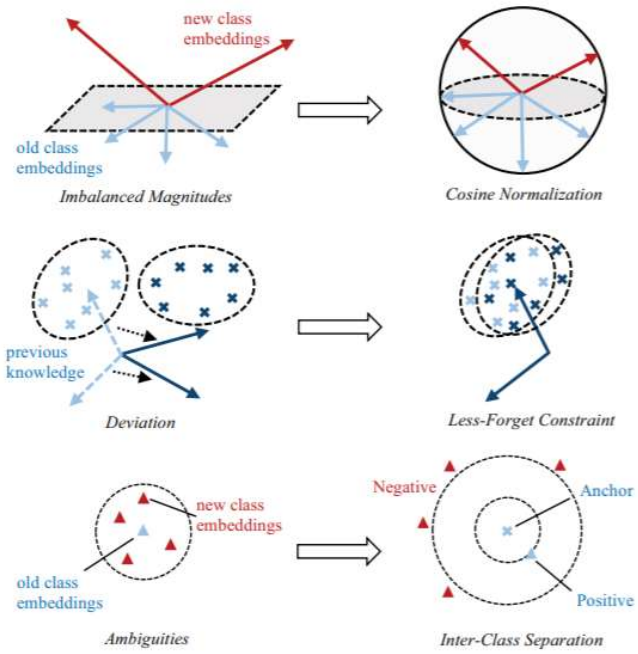

图 3 Hou 等人[7],在已有模型上提出了三种非参数化的改进来提升模型的表现（图片来自[7]）

类似的如图 3，Hou 等人[7]在类增量学习中提出了三种改进现有模型的方法，分别是在特征提取器添加一个遗忘约束，使得特征提取器能够较

好的保持原有的学习知识；在内存模型中添加了内部分段机制，使得新的类和旧的类能够尽可能大；在最后的输出层阶段添加归一化函数，使得新老类偏置能够被抵消。这三种机制一起，提升了整个模型在新测试基准上的表现，达到了 SOTA。

2.2 存在问题

我们已经介绍，增量学习的主要挑战是，是如何在不发生灾难性遗忘的情况下学习模型，对于进一步提升增量学习的学习效果，有以下问题需要重点考虑。

2.2.1 灾难性遗忘

灾难性遗忘(Catastrophic forgetting)是类增量学习面对的主要问题，类增量环境下的混合样本学习同样会陷入这类困境。以顺序方式学习任务的能力对于人工智能的发展至关重要。但是直到现在，神经网络还没有能力做到这一点，人们普遍认为，灾难性遗忘是现有神经网络结构连接模型的一个不可避免的特征。灾难性遗忘是指在一个顺序无标注的、可能随机切换的、同种任务可能长时间不再复现的任务序列中，AI 对当前任务 B 进行学习时，对先前任务 A 的知识会突然地丢失的现象。通常发生在对任务 A 很重要的神经网络的权重正好满足任务 B 的目标时。与人工神经网络形式鲜明对比的是人类和其他动物似乎能够以连续的方式学习[14]。最近的证据提示哺乳动物的大脑可能会通过大脑皮层回路来保护先前获得的知识，从而避免灾难性遗忘[14-17]。与生物大脑类似的，为了保护原有学习到的技能不被遗忘，我们需要保护参数梯度的更新，同时结合现有的训练效果。

2.2.2 模型稳定性与可塑性平衡

我们已经介绍了在灾难性遗忘的影响。与灾难性遗忘相比，灾难性干扰是在神经网络模型面对类似问题时更普遍的直接结果。灾难性干扰是指一个自然系统很难同时达成稳定性和可塑性，即所谓“稳定性-可塑性困境”。其中可塑性是指整合新知识的能力，稳定性指的是在编码新数据时保留以前的知识。这种稳定性

-可塑性的权衡是人工和生物神经智能系统的一个重要方面。相比于生物系统，人工智能系统的稳定性和可塑性还较差，其研究对于进一步的发展增量学习领域十分重要。

3．研究展望

本次文献调查中，我们考虑了类增量学习的设置和方法，在这种设置中随着数据类不断增长，模型需要不断学习，以达到在增长后的数据集上良好的分类效果。这其中在描述数据增量方面，蕴含着数据流本身是自平衡的假设。

我们知道，在自然环境中，绝大部分的数据并不是平衡的。这包括如植物的群落分布（存在某些植株特别多，某些植株特别少的情况），用户购物行为数据（日用品购买非常多，大件家具等购买非常少）。并且，在收集过程中数据是随着时间不断增长的，这表明应用增量学习的方法应该可以很自然的去处理这些问题。然而，传统的增量学习算法并不能很好的解决这个问题，长尾分布的数据会导致其在长尾类的识别准确率迅速下降，从而影响了整个模型的训练结果。

4．参考文献

[1] Yue Wu, Yinpeng Chen, et al. Large Scale Incremental Learning. arXiv preprint arXiv: 1905.13260v1 [cs.CV], 2019.

[2] 罗福星, 刘卫国. 一种朴素贝叶斯分类增量学习算法.《微计算机应用》2008年 06 期

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[4] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, Sergey Levine, et al. Online Meta-Learning. arXiv preprint arXiv:1902.08438v4 [cs.LG], 2019.7

[5] Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., and Abbeel, P. Continuous adaptation via metalearning in nonstationary and competitive environments. CoRR, abs/1710.03641, 2017.

[6] Lowrey, K., Rajeswaran, A., Kakade, S., Todorov, E., and Mordatch, I. Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control. In ICLR, 2019.

[7] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019

[8] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

[9] 耿瑞莹, 李永彬, 黎槟华. 小样本学习(Few-shot learning)综述. 阿里巴巴智能服务事业部小蜜北京团队, PaperWeekly.

[10] Santoro A, Bartunov S, Botvinick M, et al. One-shot learning with memory-augmented neural networks[J]. arXiv preprint arXiv:1605.06065, 2016

[11] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." ICML Deep Learning Workshop. Vol. 2. 2015.

[12] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in Neural Information Processing Systems. 2017.

[13] Davis Wertheimer, Bharath Hariharan. Few-Shot Learning with Localization in Realistic Settings. arXiv preprint arXiv: 1904.08502, 2019

[14] Cichon J, Gan WB (2015) Branch-specific dendritic ca2+ spikes cause persistent synaptic plasticity. Nature 520(7546):180–185

[15] Hayashi-Takagi A, et al. (2015) Labelling and optical erasure of synaptic memory traces in the motor cortex. Nature 525(7569):333–338.

[16] Yang G, Pan F, Gan WB (2009) Stably maintained dendritic spines are associated with lifelong memories. Nature 462(7275):920–924.

[17] Yang G, et al. (2014) Sleep promotes branch-specific formation of dendritic spines after learning. Science 344(6188):1173–1178.

[18] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie.The inat-uralist species classification and detection dataset.In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[19] S. Grossberg, Studies of mind and brain : neural principles of learning, perception, development, cognition, and motor control, ser.Boston studies in the philosophy of science 70.Dordrecht: Reidel, 1982.

[20] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum.2015.Human-level concept learning through probabilistic program induction.Science 350, 6266 (2015), 1332-1338.

[21] K. He, X. Zhang, S. Ren, and J. Sun.2016.Deep Residual Learning for Image Recognition.In International Conference on Computer Vision and Pattern Recognition.770-778.

[22] Zhizhong Li, Derek Hoiem. Learning without Forgetting. arXiv preprint arXiv: 1606.09282,2016.

[23] Yin X, Yu X, Sohn K, et al. Feature transfer learning for face recognition with under-represented data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 5704-5713.

[24] D. Lopez-Paz et al., "Gradient episodic memory for continual learning," in NeurIPS, 2017, pp. 6470-6479.

[25] J. Zhang, J. Zhang, S. Ghosh, D. Li, S. Tasci, L. P. Heck, H. Zhang, and C. J. Kuo, "Classincremental learning via deep model consolidation," CoRR, vol. abs/1903.07864, 2019. [Online].Available: http://arxiv.org/abs/1903.07864

二、开题报告

1. 问题提出的背景

1.1 背景介绍

我们在文献综述相关部分已经介绍了增量学习研究的意义，以及其问题设置和方法。增量学习（Incremental Learning）研究的是从数据流中学习的问题，目标是逐渐扩展所获得的知识,并将其用于未来的学习。优秀的类增量学习算法，如 iCaRL[6],LwF[22]，BIC[12]等，已经通过一些结构化的方法，如 iCaRL[8]使用内存模型来存储旧的概念向量，LwF[22]使用知识蒸馏(Knowledge Distill)来缓解新概念的学习对旧概念的梯度抵消,BIC[12]在输出层前添加偏置层学习抵消新类相对于旧类的偏置，将增量学习的分类准确率提升到了新的高度。

我们在文献综述的未来展望部分已经提出，虽然增量学习取得了一定的成功，但在数据层面来说，这些学习算法都假设数据在增量获取的时候是均衡的。对于类增量学习来说，这意味着每个类应该拥有近乎相等的样本数量。然而在现实环境中，这样的假设往往并不成立。分类问题往往关注的是自然样本的识别，其中大部分数据如植物的群落分布（存在常见的植株和罕见的植株），某个区域的动物（存在常见的猫、狗以及罕见的大熊猫等），是不均衡的。现实世界中的数据分布大多都是长尾（long-tail）（图 2）的。

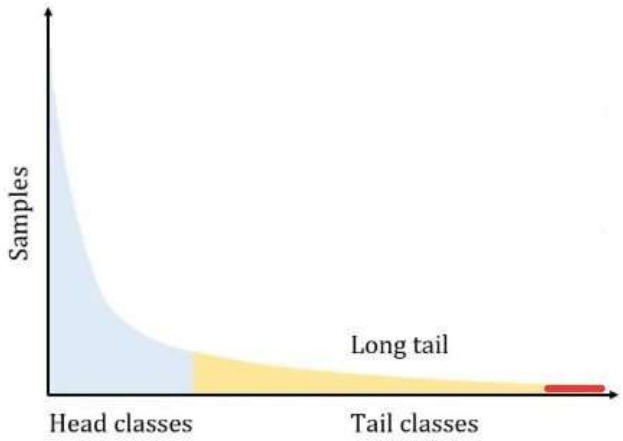

图 1 长尾(long-tail)的数据分布

如图2所示，长尾的特征是广泛存在于实际的训练数据中的，即有大量的ID（物种），它们所拥有的样本数量十分稀少，有限的样本数量不足以代表这个类别真实的分布情况，这样的ID被定义为tail class。少量的ID（物种），其样本量充分，类内多样性丰富，这样的类别定义为head class。长尾的数据分布会对分类问题造成很大的影响，Yin X[25]等人研究认为，当数据呈现长尾分布时，会导致分类器出现歧视性偏差（bias），即分类器更偏向于识别样本量充足类内多样性丰富的头部类，从而忽略了尾部类，这对尾部类而言是不公平的。在自然系统中，这也很容易理解，例如一个疾病判别系统，如果有90%的人是健康的，10%的人存在某些疾病，那么对机器来说只要将所有人都判别为健康的，它至少能够获得90%的准确率，而这个分类的准确率已经很高了。

我们认为由于尾部ID的数量庞大，而且每个尾部ID所拥有的样本数量稀少，这会导致特征空间十分混乱，大量类别的辨识度不高，使得特征空间发生扭曲，畸变。最终网络学习得到的是一个不健康的模型。当情况更糟糕时，尾部的数据可能存在大量的少样本类（Few-shot classes（1到5张样本））（图2红色部分）这意味着“尾端”ID的数据样本数量极其稀少，对于不经过特殊处理的提取器来说会导致严重的过拟合，会更进一步的影响模型效果。这种存在部分少样本类的长尾数据分布为极端情况下的长尾分布。

在现实应用中，我们认为更是如此。由于现实的大部分数据本身就是长尾分布的，我们无法提前收集大量样本再进行人工的类平衡，这样即提高了成本也不能满足模型实时学习的要求。对于人类来说，这样的数据问题不会对人类进行识别有很大的影响，通过研究在机器学习上如何能够达到人类识别水平，对于探寻类人智能发展的科学研究也是有益的。据此，相比于传统的增量学习，我们认为极端长尾分布情况下的类增量学习更具有研究的现实意义。

1.1.1 差异分析

显而易见的是，我们的研究面临着两个主要问题，分别是：1.增量学习对应的灾难性遗忘。2.长尾环境对应的尾部类歧视性偏差。这两个方向性问题分别都有了先进的解决方案，但我们通过一个简单的基准实验证明了，两种解决方法的简单结合并不能够自适应的解决本研究所面临的问题。

首先为了拟合现实环境的长尾数据分布，我们将原有的类增量数据集Cifar-

100（这是一个均衡的数据集，包含 60000 张图片，其中包含 100 个类别，每个类 600 张图片，其中训练集 500 张，测试集 100 张）通过随机指数选择的方式进行了筛选，从而将数据集改造成了指数下降的长尾分布模式，为了验证极端情况下的长尾分布，我们特意筛选了部分类别仅保留 1 到 5 个样本构成少样本类别。

为了与平衡的增量学习算法比对，我们选取了表现优异的增量学习算法 iCaRL[6]和 BIC[23]的论文原生实现进行对比测试，在进一步的长尾数据测试方面，由于 BIC 的表现要优于 iCaRL，我们只列出了 BIC 的具体准确率。实验以增量的形式进行，由于数据分为 100 个类，我们将实验分为五个阶段，每个阶段增量 20 个类别的样本，即第一阶段模型可以讲 0-20 类的样本，第二阶段模型可见 20-40 类的样本并以此类推。由于我们并没有改变每个类的顺序，同时为了公平，我们采用了每个类按照随机指数下降的方式进行样本筛选。这保证了每个类增量阶段（如 0-20 类样本阶段，20-40 类样本阶段）的数据分布是长尾的（指数下降的），同时整体的数据分布也是长尾的（指数下降的）。

Result: (/classes)	20	40	60	80	100
iCaRL*	84.40	68.30	55.10	48.52	39.83
BIC*	83.00	69.90	63.08	57.11	53.70
BIC(long-tail dataset)	64.2(3143)	56.57(3214)	48.06(3202)	39.41(3152)	32.9(3151)
BIC(long-tail dataset + Covariance pooling)	65.13(3143)	56.84(3214)	48.87(3202)	40.37(3152)	33.4(3151)

图 2 测试实验结果（第一行的数字表示当前模型一共见过的类别数量，第二到四行为实验的准确率，括号里表示的是当前增量阶段输入模型的样本数量（图片数量），对于没有括号的则为输入均衡的样本每类 500 张每个阶段 20 类一共 1000 张，符号*表示论文实现算法在均衡的类增量数据集上进行实验，long-tail dataset 表示算法在改造的长尾数据集上实验，Covariance pooling 表示协方差池化）

实验结果如图 2 所示，我们使用的 iCaRL[6]和 BIC[23]算法在文献综述部分有所介绍。在实验中 Covariance pooling 代表协方差池化，是一种特征增强方式，在 Davis 等人[2]，的论文中，它被应用于解决模型所面临的数据长尾问题，并取得了良好效果。

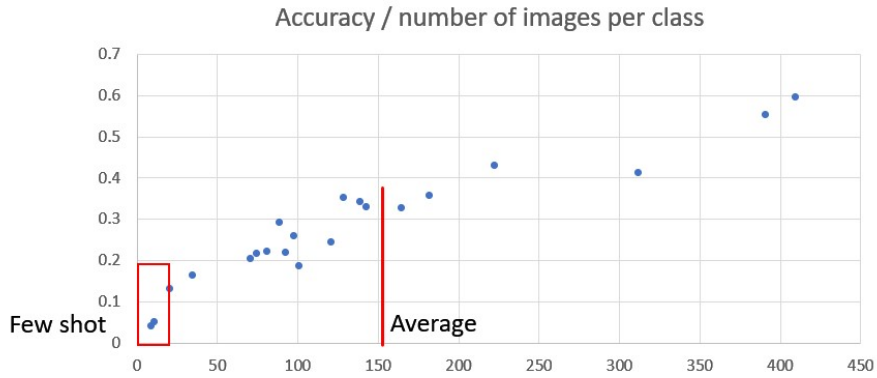

图 3 实验 BIC(long-tail dataset)中每类的样本数量对于分类准确率的影响

根据图 2 和图 3 的实验结果，我们可以从中了解到，由于长尾的数据分布，模型的识别准确率相较于均衡数据训练的模型大幅度下降。并且简单的在类增量学习算法中添加长尾数据的解决方案（Covariance pooling）并不能够有效的解决问题。根据图 3 我们可以发现，准确率下降的主要原因是模型对尾部类存在歧视性偏差，这种偏差在少样本类中表现得尤为明显（如图 3 红框所示）。

综上，本研究的问题设置与传统的增量学习和单纯的长尾数据集研究存在显著差异。同时通过实验证明了在解决方案上，简单的拼凑两者并不能够有效解决这类问题，从而证明了本研究的差异性和必要性。

1.2 本研究的意义和目的

我们在历史背景阶段已经分析了传统增量学习的局限性以及所面临的困境。进一步的，我们认为本研究的意义在于能够赋予机器在长尾甚至极端长尾的数据环境下，以增量的方式在不断地学习新数据的同时少忘记或者不忘记已有的旧经验。我们已经分析过，由于大量的现实数据是长尾甚至极端长尾分布的，这代表传统的增量学习不能够很好的处理这些数据，可能会导致对尾部少样本数据的歧视性偏差。然而增量学习的目标就在于能够长时间的分析、学习数据，这就要求数据具有一定的实时性，在这种实时性的要求下，花费大量时间人工的去平衡数据集是不实际的。因此传统的增量学习存在数据方面的局限性，我们的研究能够进一步的降低机器学习在现实中应用的数据收集成本，帮助机器学习算法更能够符合实时性的要求。

在上一节差异分析中，我们通过基准实验（图 2），得出了增量学习的方法和处理长尾数据的方法并不能够简单的混合来解决本研究面临的问题。相较于原文实现的均衡类上的实验（标*号的算法），增量学习模型在人工的极端长尾环境下准确率相较原模型大幅下降，处理长尾数据的方法也没能够很好的提升模型的表现。出现这种情况主要是因为模型会少样本类的分类准确率太差，而传统处理长尾数据分布的方法由于没有考虑到数据增量的影响，其尾部增强的方法可能同时会遭受到新的尾部类到来导致的遗忘。具体的原因需要进一步的实验分析。本研究的目的和主要挑战就在于，克服长尾甚至极端长尾环境对类增量学习算法带来的严重影响。从而为进一步的更接近现实环境的研究提供思路和实验基础。

2. 论文的主要内容和技术路线

2.1 主要研究内容

本研究主要探讨的是在长尾环境下的类增量学习算法，具体的问题设置、环境引入和差异分析等，已在前文有所介绍。根据图 3 我们已经证明了，本研究面对的主要问题是传统的长尾数据处理方法在类增量学习的情况下并不适用。接下来的研究将会从类增量学习和长尾数据结合的角度出发，考虑在新的问题环境下，为何原有的技术方法不能够通过简单的结合妥善解决本研究的问题。进一步的，我们希望能够分析出制约模型分类准确率上升的主要因素，并找出对应的方法。

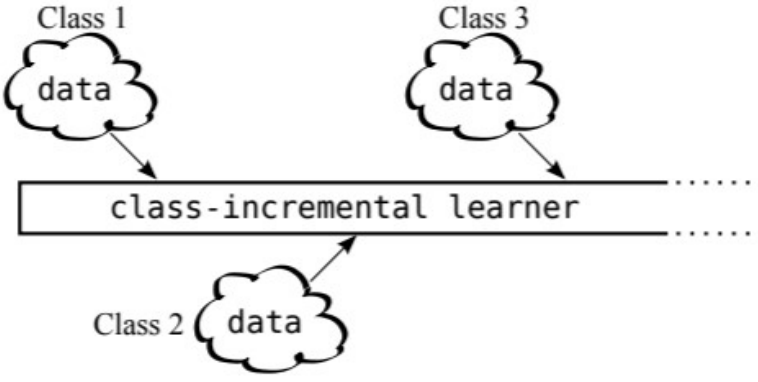

图 4 类增量学习流程(图源[6])

最后，在提出解决方法的基础上，我们希望通过一系列的基准实验证明，新

的模型或方法能够在本研究的问题基准上取得更好的分类结果。

2.2 技术路线

在数据集方面，由于缺乏现实世界的真实数据样本，与单独收集相比，我们更倾向于通过原有数据集构造人工的长尾分布。这样做的目的是既能够保证对照试验（相对于均衡类情况下的增量学习）的公平性，又能够保证数据集的质量（经过大量的实验验证）。可选的数据除了我们在前面提到的 Cifar-100 数据集以外，还可以考虑诸如 ImageNet、iNaturalist（iNat）[5]等适用于大规模物体分类的数据集。这其中 iNat 由于是一个原生收集的，“众筹”的生物群落图片集，具有充分的样本多样性，同时在数据分布上来说，它本身就是长尾的，由此它可能更适合我们的实验设置。

在技术路线方面，由于我们已经调研了大量的类增量学习方法，他们往往基于正则项和样本重放。我们在文献综述中认为，样本重放的方法往往表现更好，而正则项的方法并不适用于类间差异过大的实验。但是对于本研究来说，由于在增量过程中可以知道一个类的样本数量，对大类（样本数量多的类）来说由于信息充足特征中心较准确，基于正则项的蒸馏方式可能会对样本中心的学习造成较大影响，而对于小类（样本数据少的类）来说会出现过拟合，基于正则项的蒸馏方式一方面可以抑制小类数据的过拟合，另一方面能够减少新类数据对旧类数据梯度抵消。因此，我们可以考虑在总体上应用样本重放的方式，而对“尾端”数据（即样本数量较少的数据）进行正则项约束，从而提示模型在尾端的分类准确率。

2.3 可行性分析

本研究的研究主题是针对增量学习在长尾甚至极端长尾环境下的泛化，因此在整体架构上我们会部分的参考现有的先进的类增量学习模型，如 iCaRL[6]，BIC[23]等。类增量学习的模型框架确定了我们有可能基于重放或者正则项的方式来处理类增量学习的问题。同时由于增量数据的长尾分布，我们可能会参考一些以往的处理长尾数据分布的方法和思路。例如 Liu 等人[26]提出使用长尾数据

的头部来拟合长尾数据的尾部分布以拓宽特征空间，在控制数据分布偏差方面，Yu-Xiong Wang 等人[9]提出对长数据尾部建模的方法，获得了很好的效果。这些模型或者方法已经在原有的实验中被证明有效，虽然我们的问题是基于它们的拓展，与其并不相符，也已经经过实验证明了两者的差异。但就处理方法和思路来说，我们的问题与这些传统问题存在一定的相似性。

我们已经在技术路线中分析了一些改进措施，从理论分析角度看，这些原有方法的出发点与我们的问题环境并不冲突，我们认为可以在实验中进行进一步的测试。并且，如果本研究取得了一定的成果，可以大幅减少现实应用的数据收集压力，在现实中也具有切实的可行性。

3. 研究计划进度安排及预期目标

3.1 进度安排

本研究预期进度

2020 年 3 月 10 日~2020 年 3 月 28 日：完成文献综述、外文翻译、开题报告

2020 年 3 月 29 日~2020 年 4 月 1 日：完成前期准备，包括对照实验的代码整理，数据集的划分与准备，方向的整理

2020 年 4 月 1 日~2020 年 5 月 1 日：猜想验证和实验流程，设计对照试验，统计实验结果。

2020 年 5 月 1 日~2020 年 5 月 10 日：完成最后论文的编写，审阅，更改。

3.2 预期目标

本研究预期通过实验，分析当前研究面临的主要困难，提出创新性的模型或算法，将极端长尾分布的数据以类别增量的形式输入模型进行测试，期望能够取得相较于传统模型或算法更好的效果。在处理已有数据集如 iNaturalist、Cifar-100 等数据集成合适的混合样本数据集的基础上，使用新数据集作为测试基准，对比实验并证明在新问题上是否存在在效果更好的模型或算法。期望对数

据获取成本大的应用领域如自然物体分类、推荐系统等提供思路和方法。本研究结果预期以论文形式输出。

4. 参考文献

[1] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In CogSci, 2011.

[2] Davis Wertheimer, Bharath Hariharan. Few-Shot Learning with Localization in Realistic Settings. arXiv preprint arXiv: 1904.08502, 2019

[3] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, Sergey Levine, et al. Online Meta-Learning. arXiv preprint arXiv:1902.08438v4 [cs.LG], 2019.7

[4] 炼数之道. 增量学习综述. AI 生物医学信息学. (https://mp.weixin.qq.com/s?src=11&ligicr*pDlOYYMpiRq1i7dwp94oGFAlZXtWgGt35mDDCuNEu-&new=1). 2020.1

[5] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie.The inat-uralist species classification and detection dataset.In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[6] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

[7] Rahaf Aljund, Min Lin, Baptiste Goujaud, Yoshua Bengio. Gradient based sample selection for online continual learning. arXiv preprint arXiv: 1903.08671, 2019.3.

[8] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019

[9] Yu-Xiong Wang, Deva Ramanan, Martial Hebert. Learning to Model the Tail. NIPS 2017.

[10] Han-Jia Ye*, et al. Learning Classifier Synthesis for Generalized Few-Shot Learning. arXiv preprint arXiv: 1906.02944, 2019.6

[11] James Kirkpatrick, et al. Overcoming catastrophic forgetting in neural networks. PANS. (http://www.pnas.org/cgi/doi/10.1073/pnas.1611835114)

[12] Zhizhong Li, Derek Hoiem. Learning without Forgetting. arXiv preprint arXiv:

1606.09282,2016.

[13] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in Neural Information Processing Systems. 2017.

[14] Oleksiy Ostapenko, Mihai Puscas, et al. Learning to Remember: A Synaptic Plasticity Driven Framework for Continual Learning. CVPR2019

[15] Y. Wang, D. K. Ramanan, and M. Hebert. Learning to model the tail. In Neural Information Processing Systems, December 2017.

[16] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for fewshot learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[17] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. In International Conference on Learning Representations, 2018.

[18] C. Finn, P. Abbeel, and S. Levine. Model-agnostic metalearning for fast adaptation of deep networks. In International Conference on Machine Learning, 2017.

[19] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In IEEE Conference on Computer Vision and Pattern Recognition, June 2018.

[20] H. Edwards and A. Storkey. Towards a neural statistician. In International Conference on Learning Representations, 2017.

[21] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. Extreme clicking for efficient object annotation. 2017.

[22] H. Zhang and P. Koniusz. Power normalizing second-order similarity network for few-shot learning. In IEEE Winter Conference on Applications of Computer Vision, January 2019.

[23] R. M. French and A. Ferrara, "Modeling time perception in rats: Evidence for catastrophic interference in animal learning," in Proceedings of the 21st Annual Conference of the Cognitive Science Conference.Citeseer, 1999, pp. 173-178.

[24] Yue Wu, Yinpeng Chen, et al. Large Scale Incremental Learning. arXiv preprint arXiv: 1905.13260v1 [cs.CV], 2019.

[25] Yin X, Yu X, Sohn K, et al. Feature transfer learning for face recognition with under-represented data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 5704-5713.

[26] Jialun Liu, Yifan Sun, et al. Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective. arXiv preprint arXiv: 2002.10826 ,2020.

三、外文翻译

现实环境中带局部化的少样本学习

摘要

传统的识别方法通常需要大量的、人工平衡的训练课程，而少样本学习的方法则在小规模标注的训练数据上进行测试。与这两个极端的特例相反，现实世界中的识别问题表现为物品类别长尾状的分布模式，其中混合了粗粒度类和细粒度的类区别。我们通过实验表明，基于一个新的“meta-iNat”基准测试，以前设计的用于少量学习的方法在这些具 有挑战性的新条件下并不是现成可用的。我们引入了三个无参数的改进方法用来提高分类效果:(a)将交叉验证适应调整用于元学习从而达到的更好的训练流程，(b)在分类之前使用有限的边界框注释来定位对象的新颖的网络结构，以及(c)使用双线性池化将特征空间进行简单的无参数扩展。总的来说，这些改进使在 meta-iNat 测试上最先进模型的准确性提高了一倍，这些方法同时能够推广到以前的基准测试、更为复杂的神经结构和具有实质性的特定领域。

1．介绍

目钱，图像识别模型在 ImageNet 等基准上达到了人类的标准，但这一切的关键在于大型、平衡、有标签的训练集，每一类都有数百个例子。这一要求在许多现实场景中是不切实际的，在这些场景中，物体可能很少或者只有几个的标记训练示例。此外，获取更多标记的示例可能需要专家进行解释，因此成本太高。这一问题在应用程序(例如机器人)中更加严重，这些应用程序需要在部署时需要动态学习新概念，并且不能等待昂贵的离线数据收集过程。

这些考虑促使了对“少样本”学习问题的研究:从小样本标记集识别概念[14，18，38，41，44]。这些过去的工作建立了“学习器”的概念，它们可以根据极少数的训练例子(例如每个类别 5 个)学会区分少数从未见过的类别(通常少于 20 个)。然而，这些方法仍然面临着多重困难。

当它们被应用于现实世界的识别问题时。

首先，少样本学习方法通常假设数据集是平衡的，并且在训练过程中优化学习器，以提高精确度，通常是每个类别只有非常少的训练样本。相比之下，现实世界中的问题数据可能具有高度不平衡的长尾类分布，某些类的数据比其他类多几个数量级。因此，不管训练的例子有多少，一个实际的可用的学习器在所有类别的分类中都必须同样出色。目前还不清楚少样本学习的方法是否能够以及如果能那么如何处理这种不平衡的问题。

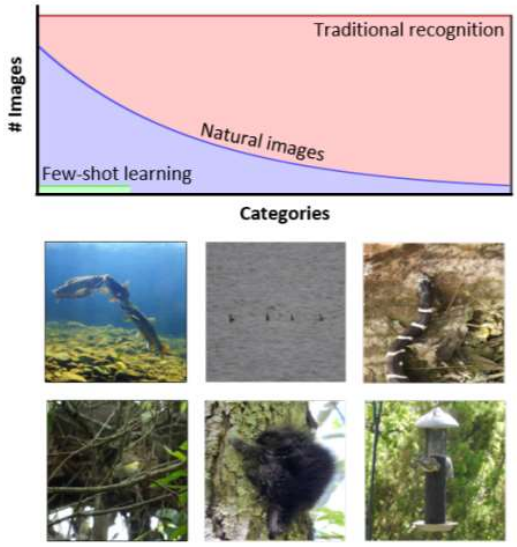

图 4 现有基准和现实世界问题之间的差异。上图:传统的识别方式使用许多均衡的巨大的数据集，而少样本测试使用均衡的数量少的类。真实世界的问题往往是困难的。底部:从左上方顺时针方向:相关物体可能重叠、微小、被遮挡、不明显(鸟在喂食器上)、模糊或很难描述[40]。

其次，少样本学习方法通常假设相关的类别数量很少，因此彼此之间有很大区别。相比之下，现实世界的应用经常涉及数以千计的仅仅有细微差别的类别。当自然图像混乱或难以解析时，这些区别尤其难以察觉(图 1，底部)。因此，学习器还必须能够对杂乱的自然图像进行精细的分类。

我们首先评估原型网络[38]，这是一种简单但最先进的少样本学习方法，基于现实的数据分布的测试基准，存在数据类长尾类分布和在“iNaturalist“数据集中[40]的一些类间差异非常细微的类。我们证明了原型网络可以在这个挑战性的基准上测试，证实了上面的直观猜想。

接下来，我们将介绍解决长尾分布、细粒度、杂乱混合识别挑战的方法。我

们引入了对原型网络的修改，在不增加模型复杂性的情况下显著提高了精度。

首先，为了解决严重的类间数据不平衡问题，我们提出了一种新的基于留一交叉验证的训练方法。这种方法使得学习器优化更加容易，学习器对数据规模的广泛分布和不均衡更有弹性。这种技术使得最终的识别精度提高了 4 个百分点。

其次，我们假设当物体很小或场景很混乱时，学习器可能会很难仅从图像级标签中识别出相关物体。为了解决这个问题，我们探索了新的学习器架构，在分类之前先定位每个主要的对象。这些学习器对标记图像的一个小子集使用包围盒标记。定位方法提高了 6 个百分点的精确度，当物体占据图像的 40%以下时提升效果更加显著。

即使在定位了对象之后，学习器也可能需要寻找物体类别之间的细微差别。现有的少样本方法仅依赖于在学习过程中来构建和形成物体的特征表示。我们证明了，直接的、无参数的调整可以显著提高性能。特别是，我们发现学习器的表征能力可以通过加杠杆的老化双线性池化（lever-aging bilinear pooling）[7, 22, 27]得到显著提高。在最初的模拟中，双线性池化显著地增加了模型参数，我们表明它可以应用于原型网络而不增加任何参数。这一修改显著提高了模型高达 9 个百分点的精确度。

总的来说，这些贡献使得原型网络和其他优秀的基础网络在我们具有挑战性的长尾基准测试上的精确度增加了一倍，而对模型复杂性的影响可以忽略不计。我们的结果表明，我们提出的方法为野外现实识别问题提供了显著优于现有技术的优势。

2. 相关工作

我们提出的技术背后的思考有着广泛的先验支持，但大多出现在不相交或不兼容的问题环境中。我们将这些概念整合到了一个统一的框架中，以便在现实场景中进行识别。

元学习（meta-learning）：以前关于少样本学习的工作主要关注于优化学习器：学习器是一个函数，它采取一个小的有标签的训练集和一个无标签的测试集作为输入，并输出对测试集的预测。这个学习器可以被表示为一个参数函数，并在一个“训练”概念的数据集上被训练，以便将其推广到新的识别领域。因为这

些方法用来训练学习器，这类方法通常被称为“元学习”。优化一般集中在将学习器的参数化上[6，15，30，37]，其更新优化的流程[29，35，36]，内置的特征提取器[12，38，41]的可推广性，或者特征空间[13，21，39，49]中的学习距离度量。一种正交（orthogonal）的方法是产生附加的合成数据[18，47]。

然而，在大多数情况下，少样本学习的分类器[6，15，21，29，35，36，38，39]仅在 Mini-ImageNet[41]或者 Omniglot[25]数据集上被评估。前者一次只呈现五个物体类别，每个类别有一个或五个训练图像。后者是一个手写十个数字的数据集，其精度通常超过 98%[12，29，30，39，41]。最近的一些工作已经大大增加了物体类别的数量，[18，44]，但仍然认为少样本的物体类别有同样数量的例子。因此，这些基准测试脱离了现实世界的条件，现实世界的条件包括困难的问题、自然的图像、许多物体种类和不同数量的训练数据[28，40，45]。许多先前的元学习方法与这些设置不兼容。值得注意的是，Wang 等人[43]设计了一种基于从普通类到少样本类的知识转移的长尾问题的方法。他们的方法与我们的改进是正交的（orthogonal）。

长尾数据集:长尾类分布在现实世界中很常见。MS-COCO[26]，the SUN database[45]，DeepFashion[28]，MINC[5]，和 Places[51]都是最常见和最不常见类别的图像数量相差多个数量级的例子。MINC 和 Places 特别值得注意，因为它们明确地被设计用来缩小数据可用性的差距[5，51]，但却表现出严重的数据类间不平衡。尽管存在这种趋势，但标准的识别基准，如 ImageNet[11]、CIFAR-10 和 CIFAR-100[23]，仍然会对其数据进行大量的整理，以确保类间保持良好的平衡并易于分离。Mini-ImageNet 和 Omniglot 少样本测试基准明确地对类间平衡进行了编码，就像其他提到的少样本测试基准[18、44、47]一样。

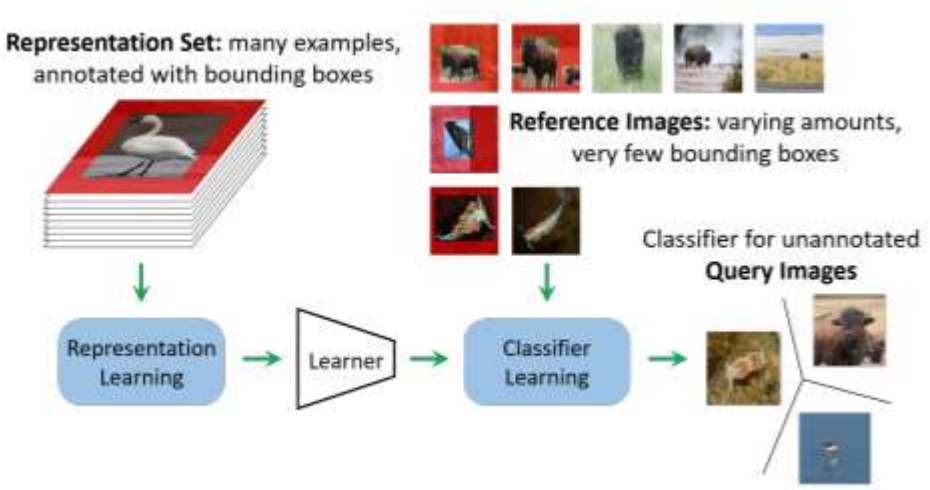

图 5　我们真实世界的学习基准。最初，许多图像都有边界框注释。然后，学习器必须适应使用不同但数 量有限的数据的新班级，并且使用很少的边界框。在野外测试时，没有注释。

改进特征空间:众所周知，特征空间的高阶扩展可以提高手工设计的特征提取器的表达能力[19，34]。最近的工作证实了类似的技术[7，22，27]的可行性，以及这些技术的可学习性的推广[8，48]，这些技术的有效近似[16，20]也提高了卷积网络的性能。这种改进在细粒度分类的问题中改善特别大，例如面部识别[4，9，27]。然而，使用得到的扩展特征空间需要参数沉重的模型，即使在少样本学习问题中[49]。我们采用双线性池化[27]作为一个真正的无参数扩展，不再有过拟合小数据集的风险。

目标定位:目标定位和识别之间存在着密切的关系。仅在图像级、基于分类的损失上训练的网络仍然可以学习对感兴趣的对象进行局部定位化[31，50]。这些学习到的定位可以作为有用的数据注释，包括用于原始识别任务[42,46,50]。然而，一些非常困难的问题可能需要昂贵的基础事实注释来开始引导标记。幸运的是，一个非常小的符号集就足以预测其余的[37]。当提供图像级类别标签时，半监督的定位方法进一步提升效果[17，24]。由于两者可以互相引导，结合识别和定位可能被证明是解决数据稀缺的特别有效的方法。

3．问题设置和基准测试

我们的目标是建立学习者，系统能够在挑战现实世界的条件下自动学习新概

念，具有重尾分布的类和微妙的类区别。每个学习者可以有可调参数或超参数。和以前的工作一样，这些参数是在概念的“表示集”([18 中的“基类”)上学习的，并有许多训练例子(见图 2)。

一旦经过训练，学习者必须归纳出一个不相交的新类别“评估集”。评估集被分成一小部分标记的“参考图像”和一大部分未标记的“查询图像”。学习者可以使用参考图像来定义新的类别集，估计这些类别的新参数(例如线性分类器)和/或微调其特征表示。

未标注的查询图像会报告最终的准确性。我们报告前 1 名和前 5 名的准确性，作为图像和评估集类别的平均值。后一种衡量标准不利于专注于大类别而忽略小类别的分类器。

对上述问题的两种解决方法可作为例证。传统的迁移学习方法是在表示集上训练一个 softmax 分类器。在评估集上，完全连接的层被具有适当数量类别的新版本替换，并在参考图像上进行微调。从测试集中查询图像。元学习方法，例如原型网络，在从表示集采样的小数据集上训练参数化学习者，教导学习者适应新的小数据集。学习者在一次通过中处理评估集，参考图像形成训练集，查询图像形成测试集。

对象位置注释:如第 1 节所讨论的，现实世界识别问题中的一个关键挑战是在混乱的场景中找到相关的对象。图像级类别标签的小集合可能不够。因此，我们为评估集中的一小部分参考图像(≤10%)提供了边界框。请注意，在[33]的实践中，通过点击极值点，这些注释很容易获得。我们对表示集进行了全面的注释，因为这样的数据集在现实世界中往往被严重地策划(图 2)。

3.1 基准测试的实现

我们现在将这个问题设置转换成一个测试基准，准确评估学习者在现实世界中的重尾问题。为此，评估集必须满足三个关键属性。首先，像在许多现实世界的问题中一样，训练集应该是严重不平衡的，稀有类和普通类之间的差别是巨大的。然而，每门课的例子数量既不能少到不必要的地步(如少于 10 个)，也不能多到不切实际的地步(如超过 200 个)。第二，与过去一次使用五个类的少量学习基准相比，[25，41]，在评估集中应该有许多(例如至少 20 个)类别，具有粗粒

度和细粒度的区别,就像在现实世界中一样。第三,图像必须具有现实的挑战性,带有杂乱和小的感兴趣区域。

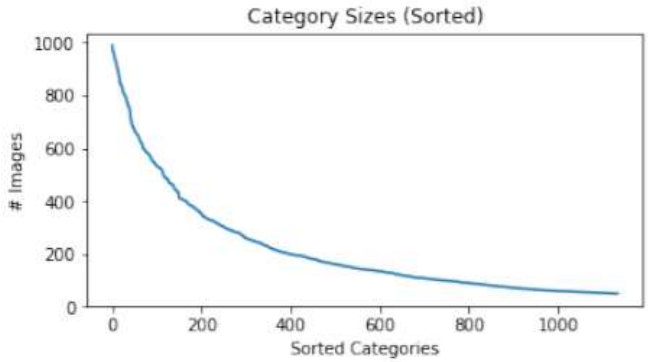

图 6 我们在评估集上运行了 10 次试验，每次试验都有不同的注释参考图像集合。

我们使用 2017 年的数据集[40]实现了我们的基准，这是一个有机收集的、众包的生物群落，具有细粒度和粗粒度的物种区分，重尾类大小分布，以及重要子集的边界框注释。在具有边界框的适当大小的类别中，80%被随机分配给表示集，其余的分配给评估集。在评估集中，20%的图像是参考图像，其余的是查询图像，总体分割为 80/4/16%的表示、参考和查询。我们提出这个“元 iNat”数据集作为元学习算法的现实的、粗尾的、细粒度的基准。Meta-iNat 包含 1，135 种动物，其分布见图 3。

虽然 meta-iNat 中的所有图像都有边界框声明，但在评估过程中只有 10%的边界框被允许使用(参见第 2-3 节)

4．改进方法

我们在原型网络[38](第 4.1 节)的基础上引入了三个轻量级和无参数的改进。批量折叠(第 4.2 节)提高训练期间的效果，并帮助学习器推广到更广泛的数据集类。少样本定位(第 4.3 节)教学习器在分类前定位一个对象。 covariance pooling(第 4.4 节)极大地提高了原型向量的表达能力，而不会影响底层网络体系结构。除了 没有参数之外，这些技术是相互兼容和相互有益的。

4.1 原型网络

我们简要回顾典型网络[38]。原型网络是一种学习者架构，旨在使用很少的训练示例来学习新颖的课程。 学习者使用特征提取器在特征空间中嵌入标记的参考和未标记的查询图像。参考图像嵌入在每个类中被 平均，以生成该类的“原型”向量。基于每个类原型的 L2 接近度对查询嵌入进行预测。

训练原型网络相当于设置特征提取器的参数，因为分类是非参数的。通过对参考和查询图像的小数据集 进行采样，在表示集上训练原型网络。这些通过网络传递，以获得查询图像的类别概率。然后最小化查 询图像上的交叉营养损失。通过这种训练，网络学习了一种特征提取器，它可以从有限的参考图像中产 生好的原型。

4.2 批量折叠

批处理折叠的动机是，在训练过程中，批处理中的每个图像要么是一个引用，要么是一个查询图像，但不是两者兼而有之。当参考图像学习形成好的类质心时，以其他贡献者为条件，查询图像被吸 引到正确的质心而远离其他。两者的梯度对于学习都是必要的，但是每个图像只有一个，所以典型的权 重更新是有噪声的。

这种引用/查询的区别还限制了网络可以处理的引用图像的数量。一个原型网络要在普通类和稀有类上工作，它必须用大量的参考图像来训练[38]。然而，增加每个批次的参考图像需要增加批次大小(这会遇到 内存限制)，或者减少查询数量(会产生更大的查询梯度)。因此，原始原型网络是为稀有类设计的。

作为一种替代方案，我们建议在每批中使用一次性交叉验证，放弃硬引用/查询分割。整个批次被视为参 考图像，并且每当它作为查询时，每个图像的贡献都从其对应的原型中减去(“折叠”)。因此，每个图像 都得到一个组合的、更清晰的渐变，既作为参考，又作为查询。此外，查询/参考图像的数量可以与批次 大小一样多/少一个。结果是在不违反内存限制的情况下，使用大的参考集进行稳定的训练。我们称这种 方法为批量折叠。

过程:设 n 为类的数量，p 为一批中每个类的图像数量。用$v_{i,\ j}$表示第 j 类

中第 i 个图像的特征向量。设 $c_j = \frac{\sum_i v_{i,j}}{p}$为第 j 类的中心。为了对 j 类中的第 i 个图像进行预测，网络使用以下类的原型:

$$c_1, c_2, \ldots, c_{j-1}, \frac{p}{p-1}(c_j - \frac{v_{i,j}}{p}), c_{j+1}, \ldots, c_n$$

开销:使用张量广播，批处理折叠可以有效地并行化。大多数机器学习库都内置了必要的广播操作，包括 NumPy[1]、PyTorch[2]和 TensorFlow[3]。

还要注意，标准原型网络预测已经包括计算每个质心和每个查询图像嵌入之间的 L2 距离。这与为每幅图 像计算(1)具有相同的渐近成本，只要查询集大小查询≈ n 的总数。一般来说，这都是事实 [38]。批量折叠的开销也往往被早期的卷积层所控制。

4.3 局部本地化

当感兴趣的对象很小并且场景混乱时，图像级标签的信息量较少，因为不清楚标签指的是图像的哪一部分。因此可以给定许多充分不同的训练图像，机器最终计算出感兴趣的区域[50]。但是但只有很少的图像和图像级标签时，区分相关特征和干扰物变得非常困难。

由于这些原因，隔离感兴趣的区域(在参考图像和查询图像上)应该会使分类明显更容易。我们考虑两种可 能的方法。在无监督定位中，学习者在代表集上开发一个与类别无关的“前景”模型。少样本定位使用 评估集上的参考图像边界框进行定位。程序:在两种方法中，定位器是一个子模块，将最终 10 × 10 特征 地图中的每个位置分类为“前景”或“背景”。该预测被计算为每个像素嵌入的负 L2 接近前景向量和背 景向量的最大值。在无监督定位中，这些向量是在表示集上优化的学习参数。在少样本定位中，定位器 得到一些用边界框标注的参考图像。我们使用这些框作为图形/背景遮罩，并对所有前景像素嵌入进行平 均，以生成前景向量。类似地计算背景向量。

定位器的输出是一个柔和的前景/背景遮罩。将要素地图与其蒙版(和逆蒙版)相乘，生成前景和背景地图， 这些地图被平均汇集，然后连接在一起。这个双倍长度的特征向量用于形成原型和执行分类。图 4 提供 了直观的解释。培训:两种本地化方法都是端到端可培训的，所以我们在分类问题中对它们进行

培训。我 们不使用额外的监管损失；定位器仅被训练用于分类。尽管如此，输出在视觉上还是相当不错的。图 5 给出了例子。

当少数几发定位器被训练成批量折叠时，在定位过程中需要额外的一轮折叠。从前景和背景向量中去除 每个图像的贡献。否则，在局部化过程中，每个图像都“看到”自己的基本事实边界框，从而阻止了对未标注图像的泛化。

4.4 协方差池化

对于困难的分类问题，可以使用诸如双线性池化[27]，Fisher 向量[34]和其他方法[4，16，22]的方法来扩展特征空间并增加 表达能力。不幸的是，传统的学习框架使用这些扩展的表示作为线性分类器、全连接的 softmax 层或 多层网络的输入，[9，20，27，49]，极大地增加了参数，并使模型倾向于灾难性的过度拟合。

然而，这些技术可以在不增加任何参数的情况下适用于原型网络。我们使用双线性池化[27]，这改进了细粒度的分类性能和概括了许多手工设计的特征描述符，如 VLAD•[19]，Fisher 矢量[34]，和视觉词汇袋(Bag-of-Visual-Words)[10]。这种方法 采用两个特征映射(例如，来自双流卷积网络)，并通过在平均汇集之前执行逐像素的外积来计算它们之间 的互协方差。在我们的定位模型中，预测的前地面和背景地图充当两个流。另一方面，我们使用要素地 图的外积。两个版本都执行带符号平方根规范化，就像在双线性池中一样，但是不投影到单位球上，因为 这严重限制了原型预测空间。

	Top-1 Accuracy		Top-5 Accuracy	
Model	Mean	Per-Class	Mean	Per-Class
Softmax	13.35±.24	6.55±.19	34.46±.30	20.05±.30
Reweighted Softmax	6.92±.19	7.88±.16	21.94±.31	22.53±.29
Resampled Softmax	1.54±.06	.99±.02	3.77±.01	2.75±.03
Transfer Learning	17.39±.24	17.61±.10	41.03±.25	40.81±.27
PN	16.07±.19	17.55±.19	42.1±.21	41.98±.18
PN+BF	20.04±.04	20.81±.08	47.86±.31	46.57±.23
PN+BF+fsL*	26.25±.05	26.29±.04	55.43±.09	53.01±.08
PN+BF+usL	28.75±.13	28.39±.15	57.90±.24	55.27±.37
PN+BF+usL+CP	32.74±.13	30.52±.13	61.32±.14	56.62±.16
PN+BF+fsL+CP*	**35.52±.05**	**31.69±.06**	**63.76±.09**	**57.33±.10**

图 7 meta-iNat 基准的测试结果，4 次试验的拥有 95%置信区间。PN 是一个原型网络，BF 是批折叠，fsL 和 usL 是少样本和无监督定位，CP 是协方差池化。*结果是 4 次试验的 10 次运行的平均值，每次运行随机抽取注释。

值得强调的是，这种扩张不会增加任何参数。与以前的型号不同，性能的所有改进都来自于功能表达能 力的提高，而不是网络容量的增加。为了强调这一区别，我们称这个版本为协方差池化。

5. 实验

我们首先在 meta-iNat 基准上展示总体结果(表 1)。我们分析本地化人员的行为，然后推广到更大的网络、 具有域转移的任务和原始的微型 ImageNet。我们使用四层卷积学习器来模拟原型网络[38]，最后加上平均池(见补充)。

5.1 meta-iNat

基准模型结果:在评估集的参考图像上从头开始训练的标准 softmax 分类器表现不佳，尤其是在稀有类上。在 训练过程中增加稀有课程只会略微提高每门课的准确性。对稀有类进行过采样会导致灾难性的过拟合。 第二个基线是转移学习:我们在表示集上训练相同的网络，但是使用类权重替换和重新训练评估集上的最 终线性层。这种方法比从头开始训练效果好得多，达到了每类 17.6%的准确率。

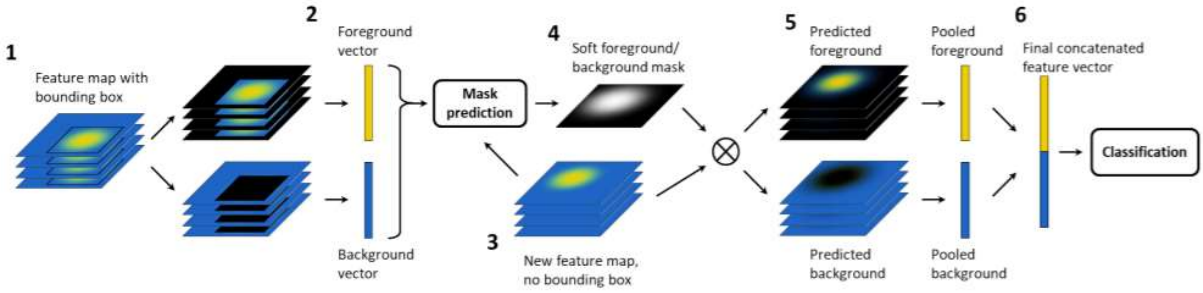

图 8 少样本定位。数据提供的边界框屏蔽了前景和背景区域(1)，它们被平均以产生前景和背景特征向量(2)。新特征图(3)上的像素特征基于与那些矢量(4)的距离被分类为前景或背景。预测掩模分离前景和背景区域(5)，它们最后被分别平均池化并拼接(6)。无监督定位学习前景/背景向量作为参数，并从(3)开始。

作为我们的第三个基线，在 meta-iNat 表示集上训练的原型网络很容易优于从零开始训练的模型，并且与迁移学习相当，而不需要任何标签重加权。这表明

原型网络在本质上是类平衡的，但在这种长尾环境下， 它并没有比迁移学习提供更多的优势。

批量折叠:经过批量折叠训练的原型网络几乎比所有基线高出 3 个百分点。图 6 中绘出了作为类别大小的 函数的每类别精度增益。我们看到了全面的收益，表明批量折叠确实提供了更高质量的梯度。同时，通 过在训练中加入更多的参考图像，批量折叠有助于模型向更大的类推广:最佳拟合线的正斜率表明大类从 批量折叠中受益更多，尽管不是以牺牲小类为代价。

局部化:引入少量局部化会使得性能获得另一个大约 6%的显著提升。请注意，10%的参考图像是 公开的，每个类别只有 1 到 20 个图像。这个相对容易的注释对性能有着巨大的影响。

图 9 少样本定位器的示例输出。最左边的图像提供了每行的前景和背景质心。网络在无监督或特殊参数的情况下学习，以隔离(大部分)选出适当的感兴趣区域。

有意思的是，无监督的本地化提供了更大的收益，大约 8 个百分点。我们假设少样本本地化表现不如它的 对手，因为它使用包围盒，一种非常粗糙的分割形式。边界框可能包含大量背景，这会破坏前景与背景 的分离。事实上，我们发现，当提供的边界框很大(例如，占据整个图像)时，少样本本地校准器无法正确定位。

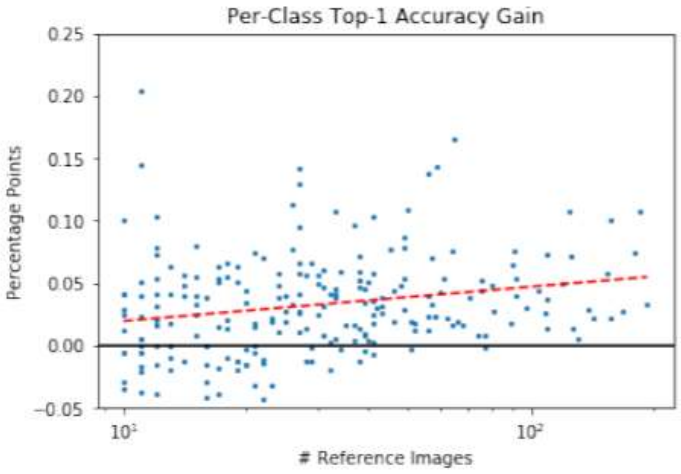

图 10 批量折叠在预期中提高了所有类大小的准确性，但对大类尤其有帮助（$r^2 = 0.05$）

正如假设的那样，当对象很小，并且边界框覆盖不到一半的图像时，定位尤其有帮助(图 7)。微小物体的 增益降低并不完全令人惊讶——当相关物体仅包含几个像素时，分类本身就更难了。协方差汇集:通过协 方差汇集，准确度再次提高，比未观测的定位提高 4 个点，比少量定位提高 9 个点。值得注意的是，协 方差池导致类平衡被打破:大类受益不成比例(图 8)。我们认为协方差空间的高维性是有意义的。小类别没有足够的参考图像来跨越空间，因此中心的作用受到影响。

无监督定位不能很好地与共方差池进行交互，这可能是因为协方差空间的维数太高，在训练过程中无法 跨越参考图像。因此，学习到的前景和背景向量可能会过度映射到表示集中的特定流形。动态计算这些 向量的少样本定位没有这个问题。我们得出结论，这两种本地化技术对于不同的设置都是有用的。

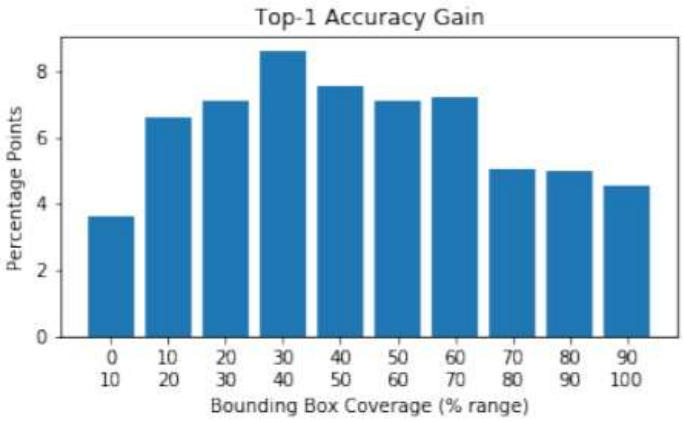

图 11 少样本学习中的局部化操作对于物体占据范围小的图片有更好的效果

使用所有这三种技术，Top-1 的准确率比基准原型网络提高了一倍。最佳表现者使用批量折叠、少样本 定位和协方差池化。消融研究作为补充提供。

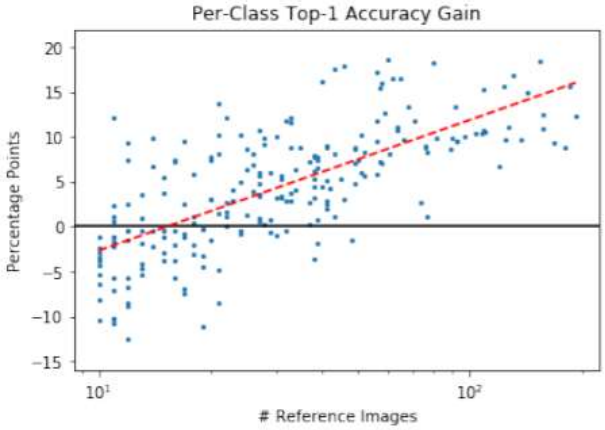

图 12 协方差池化对于图片数量多的类有较大提升，但对图片数量少的类有一定损耗（$r^2 = 0.05$）

5.2 .浅析小镜头定位器行为

接下来,我们评估良好分类准确度所需的边界框注释的数量。如表 2 所示，性能在 16%的边界框可用率 时饱和，但即使在 1%(相当于每类一个框)时，性能也仅略有下降。这种稀缺可以被放大 meta-iNat 中的类别被分成九个超级类别，所以我们也尝试每个超级类别使用一个盒子,总共九个。精确度确实显著下降，但仍比不本地化的模型要好。因此，本地化可以通过几乎不使用任何注释来获得真正的准确性，据我们所知，这是一个前所未有的发现。联合训练:尽管少发定

位器从未接受过直接训练 监督，但它仍必须与分类器联合学习。表 2 还比较了未经联合训练的定位器。对未经训练的网络应用少 量本地化会导致性能下降(“未训练”)。训练网络使用定位器，但防止通过定位器本身反向传播，也会导致性能下降(“无梯度”)。因此，本地化提供了一个有用的训练信号，但它本身必须用分类器进行训练，以获得最大的提升。

Localizer	% annotation	Mean acc.	Per-Class acc.
Untrained	10%	19.74±.03	20.42±.06
No Gradient	10%	22.77±.23	22.86±.18
Jointly trained	Supercategory	23.67±.79	24.08±.66
	1%	25.85±.11	25.96±.09
	4%	26.17±.08	26.22±.06
	16%	**26.28±.05**	**26.3±.04**
	64%	26.21±.04	26.25±.03

图 13 与基线定位器相比，随着标注的增加，Top-1 准确率上的少样本定位模型。所有型号都应用了批量折叠。

5.3 一般化

我们在三种新环境下评估我们的模型。为了测试领域转移的泛化能力，我们创建了基于超范畴的 meta-iNat 的第二个分裂。为了测试对其他网络体系结构的泛化能力，我们在 meta-iNat 上使用更强大的预处理 ResNet 体系结构来评估我们的技术。最后，使用先前文献中的评估方法，在 mini-ImageNet 上测试这些 技术。对于迷你 ImageNet 有一些预期的注意事项，我们的结果非常适用于所有设置。

超范畴元-iNat:我们希望在迁移学习更困难的环境中评估我们的结果,从表征集到评估集的转换涉及到大 量的领域转移。为此，我们构建了一个新版本的元 iNat，我们称之为超范畴元 iNat。我们不是将类别随 机分配给表示集和评估集，而是按超级类别进行划分。昆虫和蛛形纲动物(总共 354 种)构成了评估集， 其他所有动物(鸟类、鱼类、哺乳动物、爬行动物等)也是如此。)是表示集。如前所述进行培训和评估， 结果如表 3 所示。

在超范畴的 meta-iNat 上的迁移学习比在原来的环境中要困难得多。所有人的分数都一样低。然而，总体趋势 仍然实际上是一样的。批量折叠的性能优于标准原型网络，并将学习基线提升了 2 个百分点。少样本和无监督的 定位导

致相似的、基本的准确度提高(4 分)。协方差池化也提高了(5 分)，但再次导致平均精度超出每类精 度。无监督定位在使用协方差池化时执行较少样本定位，因此我们将其从未来的测试中移除。

ResNet-50:虽然批量折叠、少样本定位和协方差池化导致 meta-iNat 的实质性改进，但准确性仍然很低。 对于更强大的模型，这些改进可能会消失。为了测试这一点，我们用一个在 ImageNet 上预处理过的冻 结的 ResNet-50 来替换底层的两个原型网络层。详情可在附录中找到。结果如表 4 所示。

使用预处理的 ResNet-50 模型，可以直接从 ImageNet 向 meta-iNat 评估集执行迁移学习。冻结部分 ResNet 网络层，只训练参考图像的前两层，考虑到模型的力量，效果很差。对参考图像的整个网络进行微调效果稍微好 一些，但是降低了每类的精确度。冻结 ResNet 并把顶层训练成一个原型网络将前 1 名的准确率提高了 13 个百分点。批量折叠、少样本定位和协方差池化提供了另外 16 个点。我们得出结论，这些技术对大的 神经结构和小的神经结构都有帮助。mini-ImageNet:批量折叠、少样本定位和协方差池化提高了具有长尾类分 布的大型评估集的准确性。看看这些技术仍然有助于解决原始的、更小的少样本学习问题,我们构建了一个模拟的 mini-ImageNet 数据集，它具有相似的统计数据，但是用边界框进行了注释。我们数据集上原型网络的性能类似于已发表的数字[38]。表 5 显示了结果。

与先前结果的一个直接不同之处是，批量折叠会损害性能。批量折叠确实导致更好的训练和更低的训练损失，但是过度折叠是因为表示集更小(图 9)。

当提供五幅参考图像(“five-shot”)时，少样本定位和协方差池化做出适度但真实的改进。对单参考(“one-shot”)性能几乎没有可辨别的影响，这可能是因为参考图像很少且类很少,所以不需要更有表现力的特征 空间。尽管如此，小的改进确实表明少样本定位和协方差池化推广到少样本学习。

6. 结论

在这篇文章中，我们已经表明，过去关于类或少样本平衡基准的工作不能推广到现实的重尾分类问题。 我们表明，从有限的边界框注释中无参数定位，以及对训练和表示的改进，提供了比以前在数据丰富的 环境中观察到的更大的增

益。我们的研究只是解决更广泛的阶级平衡和数据稀缺问题的第一步。

四、外文原文

Few-Shot Learning with Localization in Realistic Settings

Abstract

Traditional recognition methods typically require large, artificially-balanced training classes, while few-shot learning methods are tested on artificially small ones. In contrast to both extremes, real world recognition problems exhibit heavy-tailed class distributions, with cluttered scenes and a mix of coarse and fine-grained class distinctions. We show that prior methods designed for few-shot learning do not work out of the box in these challenging conditions, based on a new “meta-iNat” benchmark. We introduce three parameter-free improvements: (a) better training procedures based on adapting cross-validation to metalearning, (b) novel architectures that localize objects using limited bounding box annotations before classification, and (c) simple parameter-free expansions of the feature space based on bilinear pooling. Together, these improvements double the accuracy of state-of-the-art models on meta-iNat while generalizing to prior benchmarks, complex neural architectures, and settings with substantial domain shift.

1. Introduction

Image recognition models have purportedly reached human performance on benchmarks such as ImageNet, but depend critically on large, balanced, labeled training sets with hundreds of examples per class. This requirement is impractical in many realistic scenarios, where concepts may be rare or have very few labeled training examples. Furthermore, acquiring more labeled examples