Ciliate：通过数据集与训练优化迈向更公平的基于类别的增量学习

Xuanqi Gao 西安交通大学 西安，中国
gxq2000@stu.xjtu.edu.cn

Juan Zhai 罗格斯大学 美国
juan.zhai@rutgers.edu

Shiqing Ma 罗格斯大学 美国
shiqing.ma@rutgers.edu

Chao Shen 西安交通大学 中国西安
chaoshen@mail.xjtu.edu.cn

Yufei Chen
西安交通大学
西安，中国
yfchen@sei.xjtu.edu.cn

Shiwei Wang
西安交通大学
西安，中国
shiwei.wang@stu.xjtu.edu.cn

arXiv:2304.04222v1 [cs.LG] 9 Apr 2023

摘要

由于模型存在灾难性遗忘问题，深度神经网络（DNN）需要更新，以适应新的数据分布。常见做法是采用增量学习（IL），例如通过更新输出标签进行类增量学习（CIL），利用新数据和有限数量的旧数据更新模型。这避免了使用传统方法进行从头开始的繁重训练，并通过减少需要存储的旧数据量来节省存储空间。但这也会导致公平性表现不佳。本文表明，CIL 存在数据集偏差和算法偏差问题，而现有解决方案只能部分解决这些问题。我们提出了一种名为 Ciliate 的新框架，可同时解决 CIL 中的数据集偏差和算法偏差。该框架采用一种由差异分析引导的数据集与训练优化流程，能够识别现有 CIL 方法忽略的独特且重要的样本，并促使模型从中学习。通过这一流程，根据我们对三个热门数据集和广泛使用的 ResNet 模型进行的评估，与最先进的方法 iCaRL、BiC 和 WA 相比，Ciliate 分别将 CIL 的公平性提高了 17.03%、22.46% 和 31.79%。我们的代码可在 [1] 获取。

1 引言

我们设想，搭载人工智能支持的软件的 软件 2.0时代将促成更多潜在应用（例如自动驾驶），并重塑我们的社会 [13, 35, 37]。朝着这一方向发展，以深度神经网络（DNN）为代表的机器学习（ML）技术展现出了优于其他解决方案的优势。例如，虚拟助手是基于DNN 的应用，能够理解自然语言语音指令并为用户完成任务 [2]。人工智能导航与确定性的低级控制算法（例如 PID 控制器）结合使用于自主机器人系统中 [59]。

包括深度神经网络在内的机器学习模型在大量数据上进行训练，并尝试泛化到未见过的数据。在实践中，现实世界数据的分布通常很难描述，甚至不可能描述。通过遵循严格的统计原则和最佳实践，例如平衡属于每个类别的样本数量，我们尝试生成高质量且具有代表性的训练数据集。因此，训练数据集是来自真实分布的采样数据点，无法始终准确描述现实世界的分布。当面对一组新的现实世界数据时，训练良好的机器学习模型通常需要进行更新，以适应新数据。此外，数据分布偏移在实践中也很常见。例如，自然语言处理（NLP）处理的是随着社会发展而不断变化的自然语言产物。当发生数据分布偏移时，旧模型的准确率会很低。也就是说，模型在接触新数据或数据分布发生偏移时会逐渐老化，这被称为模型灾难性遗忘现象。

传统模型训练通常需要大量训练数据集，并从随机初始化的权重开始。使用这种方法更新旧模型，需要从头记忆所有训练数据以及所有新数据（通常会进行多次更新）。这一过程耗时、存储效率低下，也不利于环境，尤其是现代模型正变得越来越庞大。例如，GPT-2 需要在 32 个 TPUv3 芯片上训练 1 周 [70]。每次有新数据更新时都重新训练这些系统，会产生大量碳足迹，且预计在未来几年还会增加 [68]。

研究人员提出了增量学习（IL）流程来缓解这一问题。例如，当模型因新数据而需要添加更多输出标签时，我们会执行类增量学习（CIL）。增量学习的基本思路是从训练良好的基础模型开始，并利用新数据更新模型。为了避免遗忘旧知识，它还会从旧的大型训练数据集中采样一小部分训练数据集。在训练过程中，它利用一个参数来平衡新旧知识，希望新模型能够泛化到旧数据分布和新数据分布。

已有研究 [41, 48, 89]以及我们的结果（见§4）均记录了增量学习（IL）方法存在模型偏差问题。通过分析，我们发现偏差问题由数据集偏差和算法偏差共同造成。具体而言，采样数据集在特征空间中可能存在偏差或不平衡。也就是说，采样数据集过度强调某些特征，或者无法涵盖所有有用且独特的特征，导致模型学习到有偏差的特征，进而作出有偏差的预测。现有基于梯度的训练算法，例如随机梯度下降（SGD），会放大数据集偏差。这些算法倾向于寻找那些容易与其他特征区分开来的特征（从梯度角度而言），而不是识别正确且稳健的特征 [5, 40]。因此，使用这些算法训练的模型无法实现理想效果，尤其是在存在数据集偏差时。现有针对增量学习公平性问题的方法开发了旨在解决偏差问题的机器学习机制。例如，iCaRL [66]

改进了采样方法，BiC [78] 处理了数据不平衡问题，而WA [89] 则通过调整权重来减轻算法偏差。一方面，这些方法忽视了借助软件工程技术观察和分析数据集偏差与模型偏差根本原因的机会，而这些分析本可为增量学习公平性问题的修正提供指导。此外，它们只针对导致增量学习偏差问题的某一个根本原因提供部分解决方案。因此，这些修正流于表面，无法全面解决问题。

受伦理意识驱动的软件工程 [8, 15, 26], 启发，我们提出了Ciliate，这是一种通过改进数据集和深度神经网络训练来修正偏差的新型增量学习框架（专注于 CIL）。Ciliate 的核心思想是遵循软件工程的最佳实践：首先观察并分析单个 IL 增量步骤中的偏差问题根因。具体而言，它对基础模型和采用传统增量学习方法训练的模型进行差异分析，以识别数据集中被忽略的样本。这些样本在训练数据集中携带独特但“次要”的特征。当被忽略时，模型会做出有偏差的预测。我们通过为每个已识别的样本标注一个分数来改进数据集，该分数反映了该样本在解决公平性问题中的重要程度。其余样本不会受到影响。为确保模型能够按预期学习此类样本，我们还改进了训练过程。对于普通样本（即重要程度较低的样本），我们沿用现有的增量学习方法，因为这些方法已经能够从此类样本中学习到足够的特征。对于标注为重要程度较高的样本，我们在每一轮训练中对其采用随机 Dropout，使不同的神经元集合能够学习这些特征。直观而言，这种改进后的训练方法促使更多神经元学习重要样本的独特特征。在进行预测时，累积效应将纠正有偏差的预测。

我们使用Python 和 PyTorch 构建了 Ciliate的原型。我们在 CIFAR-100、Flowers-102和斯坦福汽车数据集上的结果表明，Ciliate 能够有效修正偏差增量学习模型，并同时提升模型效用和公平性表现。

我们的贡献可总结如下。

- 我们提出了一个名为 Ciliate 的框架，用于解决偏差问题。在类增量学习中。它利用一种新颖的差异分析，来指导我们提出的新数据集和训练优化技术，以解决类增量学习中的数据集和算法偏差问题。
- 我们在 CIFAR-100、Flowers-102 和 StanfordCars 数据集上的评估表明，Ciliate 的表现优于最先进的方法 iCaRL、BiC 和 WA。平均而言，公平性表现指标 CWV 分别提升了 17.03%、22.46% 和 31.79%，而准确率分别提升了 11.75%、14.70% 和 1.56%（对应 iCaRL、BiC 和 WA）。

2 背景与动机

2.1 增量学习

与源代码工件一样，软件 2.0 程序的智能组件（例如深度神经网络）需要根据新数据持续更新，以支持持续集成（持续集成）和持续交付（持续交付）。例如，用于员工身份验证和识别的基于深度神经网络的人脸识别系统必须在新员工入职时进行更新。然而，更新智能组件仍然是一项具有挑战性的任务。一种直接的方法

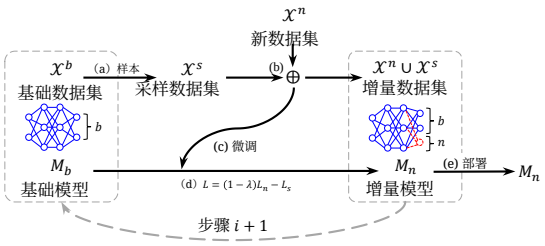


图 1：基于类别的增量学习概览。*Incr.* 是*Incremental*的缩写。

是对所有旧数据和新数据重新执行训练过程。这会造成高昂的时间和计算资源成本，尤其是在如今数据集和模型规模快速增长的情况下。此外，出于安全和隐私顾虑，在某些情况下，例如医疗保健和金融服务领域，不允许长期存储旧数据 [46, 53]。另一种传统方法是在数据偏移较小时对模型进行增量微调，也就是说，新数据与训练数据相似。但这种微调方法无法应对大量新数据。由于无法维持对已见数据的判别能力，模型可能会遭遇灾难性遗忘问题 [25, 52]。

增量学习（IL），也称为持续学习或终身学习，旨在应对从连续数据流中学习所面临的挑战。它利用数量有限的旧数据和新数据更新模型，并在所有数据上保持良好的准确率。因此，这种学习方式更具实用性。根据具体场景，IL 可分为多种类型，包括任务增量学习、基于类别的增量学习和基于域的增量学习 [34]。图 1 展示了一个简化示例，其中我们以人脸识别任务说明基于类别的增量学习（CIL）的工作方式。基础模型 M_b 在大量人脸数据 X^b上进行了训练，这些数据属于 b 个身份（即，模型有 b 个输出类别标签）。随着业务拓展，我们希望将模型扩展到 n 个更多身份及其对应的 N 份人脸数据 X^n。CIL 会从旧训练数据集 X^b中抽取并记忆一个样本数据集（或称样本）X^s，即 $\mathrm{X}^s \subset \mathrm{X}^b$，$|\mathrm{X}^s| \ll |\mathrm{X}^b|$，并使用以下损失函数在 $\mathrm{X}^n \cup \mathrm{X}^s$ 上微调基础模型 M_b：

$$L = (1-\lambda)L_n + \lambda L_s \tag{1}$$

其中，L_n 和 L_s 分别是数据集 X^n 和 X^s的交叉熵损失（通常带有温度参数），而 λ 是用于控制两种损失之间平衡的参数。简而言之，该损失将对旧数据集和新数据集的训练视为一个多任务训练过程，并使用一个超参数来平衡这两个任务。如前所述，增量学习是一个连续过程，由一系列增量步骤组成。完成一个增量步骤后，新数据集 $\mathrm{X}^n \cup \mathrm{X}^s$ 将被视为新的基础数据集，并根据超参数 λ进行采样。在后续的增量步骤中，同样的过程将在不同的数据集和标签上重复，如图 1 所示。

2.2 增量学习的公平性

在增量学习中，模型的公平性通常通过以下指标来衡量：

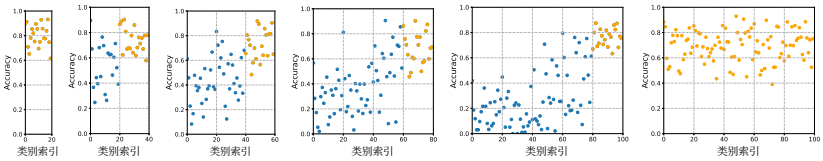


图2：每个步骤的按类别准确率。(a) 是基础模型的结果，不对应增量学习；(b)-(e) 分别是从 2nd 到 5th 增量步骤的结果；(f) 是传统非增量模型的结果，该模型使用所有类别的全部训练样本训练得到。橙色点表示新类别，蓝色点表示旧类别。各图对应模型的 CWV 分别为 0.009、0.030、0.045、0.056、0.076 和 0.015。

按类别方差 (CWV) [71]。给定一个包含 C 个类别的数据集，以及每个类别的模型准确率 c, a_c，CWV 可定义为：

$$\mathrm{CV} = \frac{1}{|C|}\sum_{c\in C}(a_c - \bar{a})^2,\ \bar{a} = \frac{1}{|C|}\sum_{c\in C} a_c \tag{2}$$

最大按类别差异（MCD） [71]。对于一个模型，给定最大和最小类别准确率$a_{\text{最大值}}$和 $a_{\text{最小值}}$，MCD 可定义为：

$$\mathrm{MCD} = a_{\max} - a_{\min} \tag{3}$$

直观而言，CWV 衡量平均差异，而 MCD 利用类别间的极端差异来估计给定模型的公平性。这些值越大，给定模型的偏差就越大。

增量学习通常存在公平性问题。为展示这一点，我们在 CIFAR-100 上开展了实验，共设置 5个增量步骤，每个步骤包含 20个类别。在每个步骤之后，我们计算每个类别的准确率，并将其绘制为散点图。我们还将公平性表现与通过传统训练获得的模型进行比较，即使用所有类别及其训练数据训练一个模型。图2展示了随着类增量学习逐步推进，准确率（每个类别）如何变化。每个点代表单个类别的准确率，橙色点代表新类别，蓝色点代表旧类别。正如我们所见，新类别的准确率显著高于旧类别，且变化幅度较大，也就是说，类增量学习模型在公平性方面表现不佳。具体而言，该模型在每个类增量学习训练步骤中的 CWV分别为 0.009、0.030、0.045、0.056和0.076。MCD值分别为 0.32、0.65、0.83、0.89 和0.87。将类增量学习模型与传统训练模型进行比较，后者的 CWV和MCD 分别达到0.015 和 0.54，我们可以看出，类增量学习模型的偏差更大。

2.3 知识蒸馏

知识蒸馏 [66, 78, 89] 是机器学习中用于将知识从复杂且计算成本高的模型（教师模型）迁移到更简单、更轻量级的模型（学生模型）的一种技术。其目标是在降低学生模型计算需求的同时，保持教师模型的性能。在知识蒸馏中，学生模型学习模仿教师模型的输出，训练目标是最小化两个模型输出之间的差异。它也常用于在增量学习 [17, 20, 89]中保留特征表示。

3 系统设计

3.1 问题陈述

在本文中，我们旨在开发一种系统，能够自动检测并修复基于类别的增量学习中的公平性问题。形式上，对于给定的基础模型 M_b 以及通过基于类别的增量学习训练得到的新模型 M_n，如果 $f(M_n) - f(M_b) > \gamma$，其中γ 是根据具体应用预先设定的阈值，f衡量模型的公平性，即 CWV 或 MCD，我们就称该模型存在公平性漏洞。我们的系统能够检测此类漏洞，并通过分析 M_n、M_b及其训练数据，训练另一个模型 $M_{\textsc{Ciliate}}$ 来修复这些漏洞。遵循先前方法 [78, 89], 中的设定，我们假设无法控制数据集 X^b、X^s和 X^n。我们的方法不会向数据集中添加新样本。我们控制训练过程，包括如何使用这些数据集中的样本以及如何训练模型。

3.2 根因分析

模型公平性通常受两个主要因素影响：数据集偏差 和算法偏差。通过分析，我们发现这两个因素都会影响在类增量学习中训练的模型的公平性。

3.2.1 数据集偏差。 数据集偏差是模型偏差的常见来源 [16, 72]。数据集会从两个方面影响模型的公平性。一方面，属于各类别的样本数量不均衡，可能导致决策不公平，使某些类别处于不利地位。举个极端例子，若模型从一百万张狗的图像和一张猫的图像中学习，它很容易忽略猫的图像（将其视为异常值），并且仍能通过将所有样本都预测为狗而达到很高的精确率。另一方面，数据质量也会影响模型的公平性。具有足够多样化特征的数据集能为模型学习提供完整的信息。否则，模型可能只会提取有限数量的特征，并基于这些不可靠的特征进行预测，从而导致公平性较低。我们的分析表明，类增量学习无法保证新训练数据的质量，进而导致模型产生偏差。

类增量学习限制了旧数据的数量，X^s。这是使用类增量学习的好处之一，但也会导致每个类别的样本数量不均衡。通常，属于新类别的数据样本会多于属于旧类别的样本，这是造成数据集偏差的根本原因之一。此外，数据集 X^s 是随机采样的。从统计学角度来看，如果样本数量不够多，采样分布会更有偏差 [40]。也就是说，当数据集 X^s 的规模不够大时，它本身就可能存在偏差。这最终会导致旧类别在特征空间中的分布范围狭窄 [66, 78]。换言之，并非所有特征都会被覆盖。

数据集 X^s。因此，旧类别的预测准确率会受到影响。如果旧特征的数量不足，模型就会因这种数据集偏差 [64] 而产生显著偏差。

首先，我们展示数据集平衡的影响。实验从一个基础模型开始，该模型采用 ResNet-32架构 [30]，，包含 20 个输出标签，每个标签使用 500个样本进行训练。我们在两个数据集上对比类增量学习：一个数据集 D_i 包含10个标签，每个标签包含 500 个样本；另一个数据集 D_j 包含相同的10 个标签，其中每个类别包含 50 个样本。随后，我们使用相同的设置进行类增量学习，即将 2000 个样本作为X^s，并使用相同的超参数集合，从而分别获得数据集D_i 和 D_j对应的模型 M_i 和 M_j。如我们所见，就训练样本数量而言，数据集 D_i 比采样数据集X^s更加平衡。这些模型的准确率、CWV 和MCD如图3所示。蓝色和橙色柱分别表示 M_i 和 M_j。如图所示， M_i 表现出比 M_j更好的公平性，证明了数据集平衡的作用。

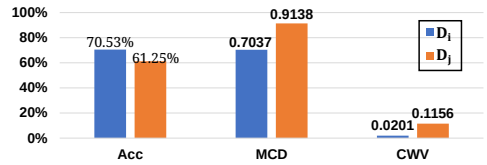
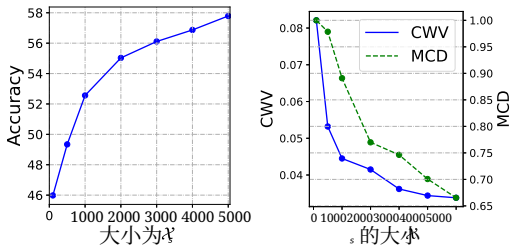


图 3：不同设置下数据集的模型性能：D_i包含 10 个标签，每个标签有 500 个样本；D_j 包含相同的 10 个标签，每个类别有 50 个样本。

我们进行的另一项实验是改变 X^s 的大小，以观察其大小对模型公平性的影响。直观来说，X^s 越大，旧数据的数量就越多，从而能在旧数据集中保留更多特征。与前一项实验类似，我们使用一个在 20 个输出标签上训练、且每个标签包含 500 个样本的 ResNet-32 模型作为基础模型 M_b。然后，我们使用不同大小的 X^s 和相同数量的新数据（即 20 个输出类别，每个类别包含 500 个样本），通过类增量学习训练新模型。我们在图 4 中报告了这些模型的准确率、CWV 和 MCD。可以看到，随着 X^s 大小的增加，模型的准确率更高，而 CWV 和 MCD 指标更低。这意味着使用更大的 X^s 数据集进行训练，能同时提升训练后模型的准确率和公平性。



(a) 准确率 **(b) CWV 和 MCD**

图 4：每个增量步骤包含 20 个类别时，在 CIFAR-100 上不同采样数据集大小下的模型平均性能X^s。

最后，为了验证特征数量与样本数量是否都会影响模型的公平性，我们还进行了另一项实验——
实验中，首先使用 500 个属于 20个类别的样本训练一个基础 ResNet-32 模型。对于 CIL 步骤，X^s的大小为 2000，新数据包含 20 个输出标签，每个标签有 500 个样本。在训练过程中，我们随机掩蔽输入像素，以模拟输入特征的损失。掩蔽比例 α分别为 0%、10% 和 20%。图 5 对比了这三个模型的准确率、CWV 和 MCD。我们分别用蓝色、橙色和灰色表示掩蔽比例 α为 0%、10% 和 20% 的情况。可以看出，随着特征损失增多，模型的准确率和公平性都会下降，这体现了特征数量的重要性。

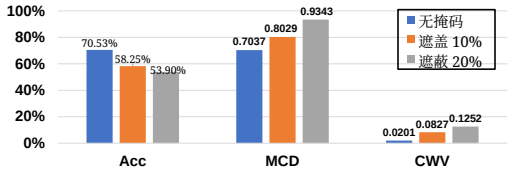


图 5：不同掩码设置下 CIFAR-100 数据集上的平均模型性能。

为探究现有方案中随机采样的数据能否提供充分的特征表示，我们在类增量学习过程中测试模型的神经元覆盖率。神经元覆盖率被广泛用于深度神经网络测试，作为测试准则来指导测试生成，以检测缺陷 [60, 79]。神经元可视为一组输入数据特征，因此神经元覆盖率能够反映模型的特征表示 [22]。在本实验中，我们从一个基础 ResNet-32 模型开始，该模型使用属于 20 个输出标签的 500 个样本进行训练。对于类增量学习步骤，X^s 的大小为 2000，新数据包含 20 个输出标签，每个标签有 500 个样本。我们以非增量学习模型的神经元覆盖率作为基准值。如果覆盖率与该值之间的差异超过 5%，则认为出现了特征表示不足问题。我们进行了五轮实验，实验结果如表 1 所示。第一列列出了实验轮次，其余列显示了每一步模型的神经元覆盖率。我们将不符合上述标准的数值加粗，可以看出，采样过程中出现了一些特征表示不足的问题。

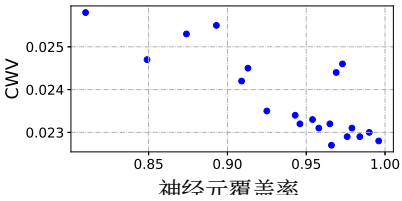


图 6：随机采样实验中按类别方差和神经元覆盖率的结果。

在本实验中，我们从一个经过训练的基础ResNet-32 模型开始，该模型使用属于20个输出标签的500个样本进行训练。对于 CIL 步骤，X^s的大小为 2000，新数据包含 20 个输出标签，且每个标签有 500 个样本。我们进行了20 轮随机采样实验，结果如图6所示。可以看出，神经元覆盖率与模型偏差之间存在负相关，这表明神经元覆盖率较低的模型无法获得足够的特征表示，从而影响模型的性能。

表 1：类增量学习过程中模型的神经元覆盖率。

步骤	1	2	3	4	5
运行 1	0.979	0.958	0.946	**0.925**	**0.810**
运行 2	0.975	0.959	0.966	0.986	0.991
运行 3	**0.943**	0.973	0.984	0.990	0.996
运行 4	0.966	0.977	0.980	0.994	0.998
运行 5	0.954	0.969	0.965	0.966	0.976
非增量					0.993

通过这些实验，我们可以得出结论：数据集偏差会显著影响通过类增量学习训练的模型的公平性。高质量的数据集应当减少数据集偏差，并强调数据特征的完整性。

3.2.2 算法偏差。模型的学习过程也可能引入偏差。首先，模型对不同样本的学习能力并不相同。现有研究 [9, 44] 表明，不同样本对深度学习模型而言具有不同的学习难度，而现有基于梯度的训练算法倾向于学习较容易的样本（梯度值更大）。具有这类重要但难以学习的特征的样本称为困难样本 [44]。一些准确率较高的模型存在严重偏差，因为模型在训练过程中优化时会舍弃难以学习的特征，因此公平性较低。对于类增量学习训练，这种影响更加明显，因为新数据集通常规模较小，针对它们的优化速度更快。因此，模型在类增量学习训练期间很难保持公平性。我们比较了常规训练的模型与被强制学习困难样本的模型。在这项实验中，我们使用与前一项实验相同的基础模型，并在包含 20 个输出标签的 CIFAR-100 数据集上训练 ResNet-32 模型，每个标签包含 500 个样本。随后，我们为一个类增量学习步骤准备一组样本，其中包含 20 个输出标签，每个标签包含 500 个样本。对于一组模型，我们采用传统的类增量学习训练对其进行训练。对于对照组，我们挑选困难样本（将在 §3.5 中讨论），并通过增加训练迭代次数以及利用 Dropout 训练更多神经元（见 §3.6），强制模型学习这些样本的特征。这两组模型的准确率、CWV 和 MCD 如图 7 所示。实线表示采用传统类增量学习训练的结果，虚线表示采用新训练方法的结果。结果表明，在类增量学习任务中，强制模型学习这些困难样本能够提高准确率和公平性。具体而言，

基于 Dropout 的训练约束使模型的准确率提升了 2.22%，CWV 提升了 8.60%，MCD 提升了 4.42%。

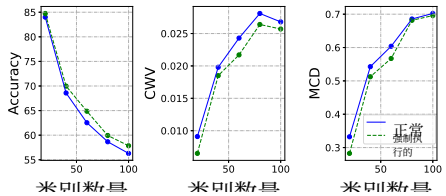


图 7：通过 Dropout训练使模型专注于学习困难样本的普通模型和模型的平均性能。

其次，数据在类增量学习中按时间顺序到达，这会导致模型的遗忘问题，也就是说，在学习后续样本时，之前的样本可能会被遗忘 [20, 33]。因此，有必要平衡模型的可塑性（学习新类别的能力）和稳定性（防止遗忘的能力）。为了证明这一点，我们比较了使用训练的模型普通交叉熵损失与采用知识蒸馏训练的模型。我们保持与上一段相同的实验设置。结果如图 8 所示。实线表示采用普通交叉熵损失训练的模型的表现，虚线表示采用知识蒸馏训练的模型。结果表明，通过利用知识蒸馏，我们在类增量学习任务中取得了更高的准确率和公平性。具体而言，采用知识蒸馏的模型在准确率、CWV和 MCD 方面分别提升了 27.14%、62.32% 和 9.47%。

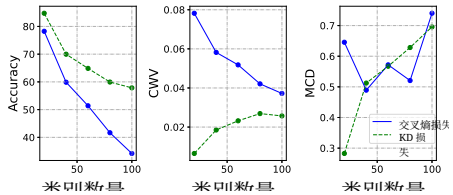


图 8：使用普通交叉熵损失训练的模型与使用知识蒸馏训练的模型的平均性能。

3.3 系统概述

基于我们的分析，我们知道，要解决模型的公平性问题，需要同时解决类增量学习中的数据集偏差和算法偏差。在 Ciliate 中，我们验证数据采样过程，以避免因特征代表不足而导致的数据集偏差，并提出了新颖的数据集精炼 和训练优化 方法，以解决数据和算法偏差。数据集精炼的目标是找到高质量的训练数据集 $\mathbf{X}^h$，其中包含类增量学习训练容易忽略的独特特征。训练优化则会修改训练流程，确保 $\mathbf{X}^h$ 中承载独特且多样化特征的少数数据能够得到充分学习，从而使模型最终实现均衡的性能。

不同于传统机器学习方法 [66, 78, 89]，，我们采用软件工程方法。图 9 展示了 Ciliate 中单个类增量学习增量步骤的整体设计。首先，我们在基础模型 M_b 上应用传统的类增量学习，得到一个新的增量模型 M_n。然后，我们对两个模型的输出进行差异分析，并利用分析结果筛选样本，从而将训练数据分为两部分：$\mathbf{X}^h$ 和 $\mathbf{X}^l$。最后，我们对这两部分数据采用不同的训练方法，得到优化后的模型 M_c，作为此增量步骤的输出。

Ciliate的整体算法如算法1所示，记为过程 Ciliate。它以基础模型 M_b、有偏差的类增量学习模型 M_n和训练数据 $\mathbf{X}^t$ as输入，并输出一个修正后的模型。首先，Ciliate 对 M_b 和 M_n 的模型输出执行差异分析（第1-6 行）。其基本思想是分析类增量学习前后模型输出的变化，并评估各个输入对两个模型 M_b 和 M_n的重要性。由于传统的类增量学习对每个样本使用相同数量的迭代次数进行训练，少数独有特征很容易被忽略，从而导致偏差。通过分析模型输出的差异，我们可以估计每个输入样本所携带的独有特征，并识别出在旧的类增量学习中消失的特征。基于这些结果，Ciliate 执行数据集精炼，突出对公平性再训练而言重要的样本，记为$\mathbf{X}^h$ （第 7行）。其余数据集记为 $\mathbf{X}^l$。最后，Ciliate 执行精炼训练，促使有缺陷的模型学习无偏差特征（第 8-12 行）。对于那些

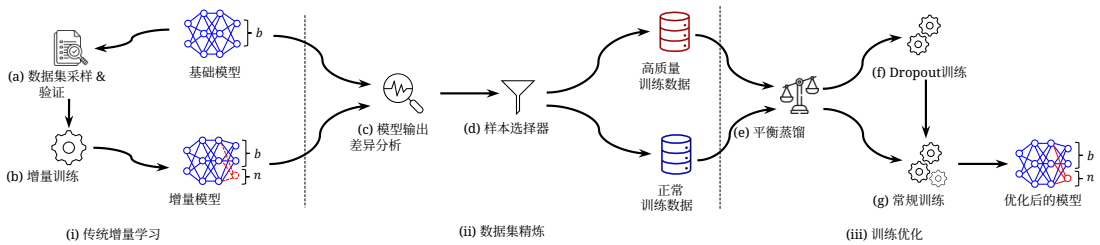


图9：Ciliate 的一个增量步骤。

算法1CILIATE 算法

输入： M_b：增量训练前的基础模型
输入： M_n：通过增量训练获得的有偏增量模型
输入： $\mathcal{X}^t$：训练数据集
输出： M_c：修复后的优化后的模型

1: **过程**CILIATE
2: $L_d \leftarrow [\,]$
3: **for** $s \in \mathcal{X}^t$ **do**
4: $P.div \leftarrow JSLoss(M_b(s)\ M_n(s))$
5: $P.sample \leftarrow s$
6: $Append(PathList, P)$
7: $\mathrm{X}^l, \mathrm{X}^h \leftarrow GetRefined(L_d, \mathrm{X}^t, \eta)$
8: $M_c \leftarrow M_b$
9: $E \leftarrow GetErrorset(\mathrm{X}^t, M_n)$
10: **for** $h \in \mathcal{X}^h$ **do**
11: $DropoutTraining(M_c, s)$
12: **对于** $l \in \mathcal{X}^l$ **执行**
13: $OrdinaryTraining(M_c, l)$
return M_c

输入： L_d：与训练样本对应的散度列表
输入： $\mathcal{X}^t$：训练数据集
输入： η：用于控制比率的超参数
输出： $\mathcal{X}^l$，$\mathcal{X}^h$：分别表示普通质量和高质量数据集。
14: **过程**GetRefined
15: $L_s \leftarrow Sort(L_d.div)$
16: **对于** $i \leftarrow 0\ to\ \eta \times len(L_s)$ **执行**
17: $Append(\mathrm{X}^h, L_s.sample)$
18: **对于** $i \leftarrow \eta \times len(L_s)\ \ to\ len(L_s)$ **执行**
19: $Append(\mathrm{X}^l,\ \ L_s.sample)$
返回 $\mathcal{X}^l$，$\mathcal{X}^h$

对于 X^l 中的样本，我们使用常规训练方法；而对于 X^h，我们进行 Dropout训练，促使有缺陷的模型考虑更多特征，以避免产生缺陷。通过这种精细化训练，Ciliate 促使模型学习更多无偏特征，而非偏差特征，从而缓解公平性问题。

3.4 数据采样

在 CIL 流水线开始时，我们沿用传统方法对基础数据集进行采样，以获得用于增量训练的新数据集（图 1(a)）。如 §3.2 所示，采样数据可能存在代表性不足的问题，从而降低训练效果和模型的公平性表现。因此，我们增加了一种由神经元覆盖率引导的重新采样机制。更正式地说，给定模型的神经元 M、$|M| = n_1,\ \ n_2,\ \ .....$以及输入数据 $X = x_1,\ \ x_2,\ \ .....$，对于给定输入 x，我们用 $n(x)$ 表示神经元 n 的输出值。对于激活阈值 t，神经元覆盖率定义如下：

$$NC(M, \mathcal{X}) = \frac{|n|\forall x \in \mathcal{X}, n(x) > t|}{|M|} \tag{4}$$

在 Ciliate 中，我们将采样数据集 X^s输入基础模型 M_b，并计算 M_b 的神经元覆盖率。如果 $NC(M_b,\ \mathrm{X}^s) > \beta$，我们则认为该数据集采样充分，并直接将其用于增量训练。否则，我们会重新采样该数据集。我们研究参数 t 和 β 的影响，并在 §4.4 中给出结果。

3.5 数据集精炼

3.5.1 模型输出差异分析。 模型输出差异分析主要旨在帮助 Ciliate 理解模型中的公平性缺陷，即哪些样本对模型的公平性表现至关重要。首先，我们将输入样本送入模型 M_b 和 M_n，并获取其输出值，分别记为 $M_b(s)$ 和 $M_n(s)$。然后，Ciliate 计算模型对每个样本的预测的 Jensen-Shannon 散度值（算法1第 4 行）。最后，我们将该散度值与样本关联，并将其存入列表（记为 L_d），供后续步骤使用（第 5–6 行）。

正如我们在§3.2.1中所讨论的，模型偏差主要由训练数据集引入。训练前后模型输出之间的散度可以反映输入样本对基础模型 M_b训练的影响程度。也就是说，它反映了输入样本对模型公平性影响的重要程度。因此，我们迭代计算每个输入样本的散度值，以获得一个能够反映样本重要性的标准。在数学上，有许多计算散度的方法，包括库尔贝克-莱布勒散度 [43], 、詹森-香农散度 [50], 和海林格散度 [31]。我们测试了它们的性能，最终选择詹森-香农散度作为标准（更详细的讨论请参见 §4.3）。分布 P 与分布 Q 之间詹森-香农散度的计算公式如下：

$$JS(P, Q) = \frac{1}{2}\sum P \log(\frac{P}{M}) + \frac{1}{2}\sum Q \log(\frac{Q}{M}) \tag{5}$$

其中 $M = \frac{1}{2}(P + Q)$。获得散度值后，我们将其与样本逐一匹配，以便进行下一步：样本选择。

*3.5.2*样本选择。样本选择的目标是选择重要样本，并将其整合到数据集 X^h中。算法1中的过程 *GetRefined* 展示了样本选择过程，该过程将旧训练数据集 X^b 分为两个数据集X^h和 X^l。首先，我们根据上文获得的 L_d 的发散值对其进行排序，以得到有序列表 L_s（第15行）。然后，我们将 L_s 中对应于较大值部分的样本添加到 X^h中，其余样本添加到X^l 中（第16—19行）。截断点由预设超参数 η 与 L_s长度的乘积决定，即 $len(L_s)$。具体而言，对于按降序排序的列表 L_s

顺序，我们通过以下公式计算截止点：

$$P_c = \eta \times len(L_s) \tag{6}$$

我们将索引低于截断点的任何样本指定为重要样本，并将其添加到X^h中。对于索引大于截断点的样本，我们将其添加到X^l中。

如前所述，每个样本的重要性由 L_d反映。通过对 L_d进行排序，样本的重要性得以自然排名，从而可以区分出最重要的样本。我们可以通过调整超参数 η来改变高质量数据集 X^h 的大小，这会改变模型在后续阶段 训练优化 中学习数据特征的尺度。

3.6 训练优化

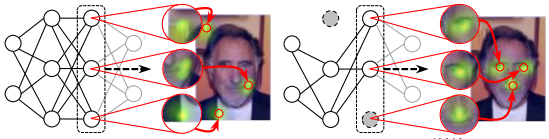
3.6.1 平衡蒸馏。 训练优化旨在帮助Ciliate 缓解由模型学习和遗忘引入的算法偏差（详见§3.2.2）。受PODNet [20] 和 JTT [49], 的启发，我们提出了一种改进的损失函数，即平衡蒸馏。正如我们在 §3.2.2 中所讨论的，适应新数据的模型往往会遗忘之前学到的知识。为此，我们通过蒸馏损失 L_r引入了严格约束，帮助模型保持先前的知识 [32]。对于偏差增量模型 M_n中那些分类错误的样本 E，我们将其放入错误集，并使用交叉熵损失 L_{CE}对其进行训练（算法1第9行）。对于分类正确的样本，则使用 L_r进行训练，以保留这些正确的知识。因此，错误集 E可以形式化表示如下：

$$E = \{(x,y)|(x,y) \in X^t \wedge M_n(x) \neq y\} \tag{7}$$

因此，总损失函数的公式如下：

$$L = \lambda \sum_{(x,y)\in E} L_r(x) + \sum_{(x,y)\notin E} L_{CE}(x) \tag{8}$$

我们比较了使用和平衡蒸馏损失训练所获得的结果，详见 §4.3。



(a) 常规训练 (b) Dropout训练

图 10：Dropout 技术示例。

*3.6.2Dropout*训练。 首先，我们将基础模型 M_b 作为训练起点（算法1第8行）。然后，我们在 X^h上进行 Dropout训练。最后，我们进行常规训练 X^l。请注意，这两个训练过程都各自包含多个训练步骤周期。

训练优化主要在 X^h 数据集上进行训练时应用丢弃技术，以帮助模型学习更加公平的表征并提升其公平性。Dropout [69] 是一种正则化技术，在每次训练迭代期间，神经网络中的一定比例节点会被随机“丢弃”（即设为零）。这会迫使其余节点学习更加稳健且相互独立的特征，并有助于防止过拟合，即模型变得过于复杂，只记忆训练数据，而不是从中学习。在测试期间，会使用所有节点，但它们的输出会根据它们在训练过程中得以保留，从而使Dropout的整体效果仍然存在。Dropout是深度学习中广泛使用的一种技术，已被证明能够有效提升神经网络在包括缓解公平性 [27, 76]在内的广泛任务上的泛化性能。Dropout训练迫使神经网络中的每个神经元都有机会学习每个样本的不同表示，从而使困难样本更容易被模型记忆和学习，进而减轻算法偏差。

我们还比较了不使用 Dropout 进行训练或完全使用 Dropout 进行训练所获得的结果，详见 §4.3。

4 评估

我们旨在通过实验回答以下研究问题：

研究问题1： Ciliate 能在多大程度上修正类增量学习中的偏差模型?

RQ2: 数据集精炼和训练优化如何影响 Ciliate 的性能?

问题3： 在Ciliate 中，重要的可调参数/函数有哪些，它们有什么影响?

研究问题4： 有哪些重要经验可以帮助开发更好的增量学习流程?

研究问题5： Ciliate 在修正类增量学习中的偏差模型方面有多高效?

4.1 设置

4.1.1 硬件和软件。我们在一台服务器上进行实验，该服务器配备 64 核 Intel Xeon 2.90GHz CPU、256 GB 内存和 4 块 NVIDIA 3090 GPU，运行 Ubuntu 16.04 操作系统。

4.1.2 数据集。我们在三个常用数据集上评估这些方法：CIFAR-100、Flowers-102 [54], 和 Cars [42]。

- **CIFAR-100：** 包含 6 万张 32×32 RGB 图像，分属 100 个类别。每个类别有 500 张训练图像和 100 张测试图像。2000 个样本被保存为样本代表。
- **Flowers-102：** 包含 8189 张 RGB 图像，分属 102 个类别。每个类别包含 40 至 258 张图像。训练集和验证集各包含每个类别 10 张图像（共 1020 张）。测试集包含剩余的 6149 张图像。1000 个样本被保存为样本代表。
- **Cars：** 包含 16,185 张图像，分属 196 个汽车类别。该数据集分为 8,144 张训练图像和 8,041 张测试图像，每个类别大致按 50:50 的比例划分。1000 个样本被保存为样本代表。

4.1.3 模型。为与基线方法进行公平比较，我们复现了基线方法，使用相同的模型，并保持相同的设置。我们使用 SGD 训练一个 32 层的 ResNet，并将批量大小设为 32。学习率初始为 0.1，并在第 100、150 和 200 个 epoch 后降至前一学习率的 1/10（总计 250 个 epoch）。我们采用随机裁剪、水平翻转和归一化进行数据增强。

4.1.4 基线方法。我们将 Ciliate 与其他最先进的基于类别的增量学习方法进行比较，包括 iCaRL、BiC 和 WA。

iCaRL。作为最早将采样引入类增量学习（CIL）的方法之一，iCaRL 提出根据数据特征空间表示对旧数据进行采样 [66]。该方法会提取所有样本的表示，然后计算每个类别的平均表示，并通过迭代为每个类别选择样本代表。在每一步中，都会选择一个样本代表，使其加入某个类别的样本代表集合后，更新后的样本代表质心与真实类别质心之间的距离最近。

BiC. Wu 等人 [78] 提出了一种偏差校正方法，以缓解增量学习中的不平衡问题。BiC 会在网络中添加一个专门用于校正任务偏差的额外层。随后，BiC 将一个训练阶段分为两个阶段。在第一阶段，BiC 按照常规类增量学习的方式训练模型。在第二阶段，BiC 冻结模型中的所有参数，并从少量训练数据中划分出一部分作为验证集，以学习所添加的偏差校正层。

WA. Zhao 等人 [89] 提出了 WA，该方法通过添加新层来对齐模型权重，从而提升类增量学习的性能。在常规类增量学习训练之后，他们会对全连接层中的偏差权重进行调整，帮助模型修正存在偏差的权重。

这些工作从机器学习的角度提升了 CIL 的性能，并增强了 CIL 的通用性。然而，它们都引发了公平性问题。原因在于，它们都没有约束采样数据集本身的质量，也没有使用任何指标作为指导。因此，实际上，这些方法是在有偏差的数据集上进行 CIL 训练，这使模型难以实现较高的公平性。

4.1.5 评估指标。 我们使用三项指标将 Ciliate 与基线方法进行比较：准确率、CWV 和 MCD（见 §2.1）。

4.2 研究问题1：修复性能

实验设计： 为了评估 Ciliate 的修复性能，我们测试以下模型：朴素训练（非增量）模型，以及由 BiC、WA、iCaRL 和 Ciliate 训练的类别增量学习模型。朴素训练模型使用完整数据集进行训练，其训练器设置与类别增量学习模型相同，具体如 §4.1 所述。对于类别增量学习模型，训练从基础模型（步骤 1）开始；由于我们采用 5 步类别增量学习，因此初始使用数据集总体标签数量的五分之一。随后，在每个增量步骤中，我们使用一个由增量数据集（包含数据集总体标签数量的五分之一）和采样数据集组成的数据集来训练模型；采样数据集根据上一步的训练数据集进行更新，同时保持相同的总体大小。每个增量步骤完成后，我们在测试数据集的一个子集上评估更新后的模型，该子集包含与已见标签相关的样本。每个数据集的采样数据集大小见 §4.1。我们从效用和公平性两个方面比较 Ciliate 与其他算法的性能。

结果： 比较结果呈现在表2中。第一列列出了三个数据集。第二列展示了不同的算法。其余各列列出了每个步骤中的模型性能，包括准确率（准确率）、按类别方差（CWV）和最大按类别差异（MCD）。最佳结果以粗体突出显示。实验结果证明了我们的算法在修复模型公平性问题方面的有效性。首先，Ciliate 可以有效缓解基于类别的增量学习中的公平性偏差。其次，Ciliate 实现了最高的效用在所有算法中。表2分别展示了增量学习在 CIFAR-100、花卉和 Cars 上的公平性提升。Ciliate 在最终增量准确率和平均增量准确率方面均优于最先进的方法。我们还将模型性能与非增量模型进行了比较。

对于公平性表现，Ciliate 在所有模型中取得了最高的 CWV，分别在 CIFAR-100、花卉和 Cars 上比 iCaRL 高 33.26%、3.77% 和 14.07%；比 WA 高 56.75%、22.48% 和 16.13%；比 BiC 高 47.94%、11.49% 和 7.94%。在 Cars 数据集上，Ciliate 甚至比非增量模型高出 3.35%。在 MCD 方面，在 CIFAR-100 上，Ciliate 比 iCaRL 高 33.26%，比 WA 高 4.17%，比 BiC 高 3.89%。对于花卉和 Cars 数据集，它们在 MCD 表现方面没有显著差异。我们发现，这主要是因为这两个数据集的分布更难学习，且旧数据的采样规模较小，导致始终有一个或某些类别无法得到全面学习。此外，表2表明，Ciliate 通过成功解决类别增量学习模型的公平性问题，在提升准确率方面具有优势。Ciliate 的平均效用优于所有类别增量学习模型。与非增量模型相比，在 CIFAR-100、花卉和 Cars 上，Ciliate 的准确率分别下降了 19.75%、25.18% 和 16.39%；iCaRL 分别下降了 21.98%、35.96% 和 27.55%；WA 分别下降了 20.36%、26.84% 和 17.67%；BiC 分别下降了 27.86%、38.64% 和 24.42%。平均而言，与最先进的方法 iCaRL、BiC 和 WA 相比，Ciliate 的 CWV 分别提高了 17.03%、22.46% 和 31.79%。

分析： 从表2中，我们得出几点观察。首先，与其他三种类增量学习方法相比，Ciliate 的效果更好（测试准确率更高），公平性也更好（公平性表现更优）。与此同时，Ciliate 修复公平性缺陷后，模型准确率也有所提升，这表明 Ciliate 不会降低整个模型的有效性。因此，我们可以说，Ciliate 能有效修复类增量学习模型中的偏差，同时提升模型效用和公平性表现。

4.3 RQ2：优化的影响

实验设计： 本节旨在展示 Ciliate 两个核心阶段——数据集精炼和训练优化——的有效性。为此，1）对于数据集精炼，我们比较 Ciliate、未进行验证的 Ciliate，以及未进行样本选择的 Ciliate 的修复性能；后者即随机数据集精炼方法，其中随机获得的精炼数据集大小与 Ciliate 的相同；2）对于训练优化，我们比较 Ciliate、未采用平衡蒸馏的 Ciliate，以及未采用选择性训练的 Ciliate（包括纯 Dropout 和纯普通方法）的修复性能。我们遵循 §4.2 中的类增量学习设置，并在 CIFAR-100 数据集上开展实验。我们从平均表现（不包括基础模型）和最后一步获得的模型表现两方面比较这些方法。

结果： 比较结果的详细信息见表 3。第一列列出各方法。接下来的三列显示除第一步外所有增量步骤的平均结果，剩余列则显示最后一步的模型性能。所有增量步骤的详细结果见补充材料。总体而言，在所比较的数据集和训练技术中，Ciliate 实现了最佳的修复性能。

表2：三个数据集上的结果。最佳结果以粗体显示。

数据集	增量步骤 / 方法	1 Acc	1 精确率	1 召回率	1 CWV	1 MCD	2 Acc	2 精确率	2 召回率	2 CWV	2 MCD	3 Acc	3 精确率	3 召回率	3 CWV	3 MCD	4 Acc	4 精确率	4 召回率	4 CWV	4 MCD	5 Acc	5 精确率	5 召回率	5 CWV	5 MCD
CIFAR-100	iCaRL	81.95	79.83	79.77	0.0750	0.4173	69.85	68.13	68.09	0.0688	0.5667	63.07	61.56	61.49	0.0551	0.6750	57.30	55.87	54.82	0.0461	0.7160	55.03	52.89	52.34	0.0415	0.7699
	WA	**85.30**	82.50	82.46	0.0726	0.3788	**71.73**	70.68	70.26	0.0837	0.5385	**64.97**	63.95	63.82	0.0784	0.6814	59.58	57.22	57.30	0.0699	**0.6403**	56.17	53.21	53.77	0.0641	0.7365
	BiC	84.95	82.70	82.66	**0.0691**	0.3833	68.30	66.98	66.67	0.0740	0.6322	60.13	59.08	58.99	0.0657	0.6618	55.70	53.47	53.49	0.0619	0.7589	50.88	48.75	48.37	0.0532	0.7343
	Ciliate	84.30	82.18	82.17	0.3214	**0.3178**	70.38	67.92	66.25	**0.0177**	**0.5141**	63.92	61.84	60.50	**0.0216**	**0.5821**	**59.86**	57.77	56.58	**0.0262**	0.6453	**56.60**	54.08	53.97	**0.0256**	**0.6888**
	非增量	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.53	67.21	66.33	0.0201	0.7037
花卉	iCaRL	90.42	91.01	88.70	0.0969	0.8000	61.26	59.00	57.80	0.0985	1.0000	59.90	58.77	58.91	**0.0877**	1.0000	62.09	61.80	59.04	0.1307	1.0000	57.88	57.74	55.25	0.1146	1.0000
	WA	**95.81**	95.80	93.56	0.0865	0.8000	**72.84**	73.88	69.56	0.1289	1.0000	68.67	68.93	64.02	0.1417	1.0000	69.52	69.96	63.87	0.1505	1.0000	66.13	66.41	62.14	0.1423	1.0000
	BiC	93.75	93.69	91.77	0.0853	**0.7500**	62.89	63.92	60.01	0.1474	1.0000	61.02	61.37	59.98	0.1768	1.0000	60.72	61.13	59.73	0.1431	1.0000	55.46	55.82	53.66	0.1246	1.0000
	Ciliate	94.84	94.76	92.94	**0.0662**	**0.7500**	69.47	72.05	64.66	**0.1058**	1.0000	**70.36**	71.43	67.27	0.1044	1.0000	**71.11**	71.57	66.48	**0.1211**	1.0000	**67.63**	67.92	64.85	**0.1103**	1.0000
	非增量	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	90.38	90.21	88.33	0.0933	0.7500
Cars	iCaRL	84.94	84.09	83.87	0.1581	1.0000	70.19	72.18	69.56	0.1579	1.0000	62.73	64.09	62.41	0.1374	1.0000	56.74	62.92	54.81	0.1349	1.0000	53.31	65.93	53.17	0.1274	1.0000
	WA	87.98	87.24	87.00	0.1203	0.7228	**74.67**	74.33	73.80	0.1449	1.0000	64.71	72.39	63.74	0.1400	1.0000	60.15	70.78	59.83	0.1326	1.0000	60.59	72.60	60.05	0.1306	1.0000
	BiC	88.08	87.48	87.10	**0.1130**	0.8834	73.54	72.61	71.77	0.1399	1.0000	64.25	69.03	63.62	0.1361	1.0000	59.40	68.85	58.79	0.1244	1.0000	55.61	69.02	55.07	0.1189	1.0000
	Ciliate	**88.24**	87.61	87.06	0.1213	**0.6667**	72.30	71.92	70.66	**0.1250**	1.0000	**65.60**	76.78	64.14	**0.1165**	1.0000	**61.23**	75.74	59.94	**0.1035**	1.0000	**63.12**	76.94	60.44	**0.1095**	1.0000
	非增量	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	73.59	73.66	71.25	0.1133	0.8077
ImageNet	iCaRL	84.74	82.63	82.47	0.0232	0.6014	67.46	66.35	65.80	0.0217	0.6487	53.97	53.14	52.42	0.0197	0.6833	46.22	45.31	44.42	0.0209	0.7052	42.75	42.11	41.25	0.0183	0.7471
	WA	91.17	89.87	89.03	0.0227	0.5735	85.34	84.22	83.94	0.0214	0.5990	80.83	79.31	78.68	0.0222	0.6419	76.68	74.77	74.05	0.0213	0.6637	70.29	69.52	69.16	0.0204	0.6915
	BiC	89.17	87.75	87.39	0.0215	0.5749	84.35	81.97	81.82	0.0199	0.6062	78.09	77.14	76.86	0.0194	0.6318	73.67	71.79	71.28	0.0186	0.6721	66.50	65.89	65.13	0.0178	0.6825
	Ciliate	88.86	87.94	87.42	0.0322	0.5311	84.45	82.17	81.73	0.0246	0.5571	80.53	79.15	78.48	0.0215	0.5931	76.47	75.16	75.25	0.0189	0.6152	73.79	72.92	72.57	0.0181	0.6478
	非增量	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	75.37	75.14	74.98	0.0109	0.6953

表 3：由不同组件训练的模型之间的比较。报告了平均性能（基础模型除外）以及最后一步获得的模型的性能。

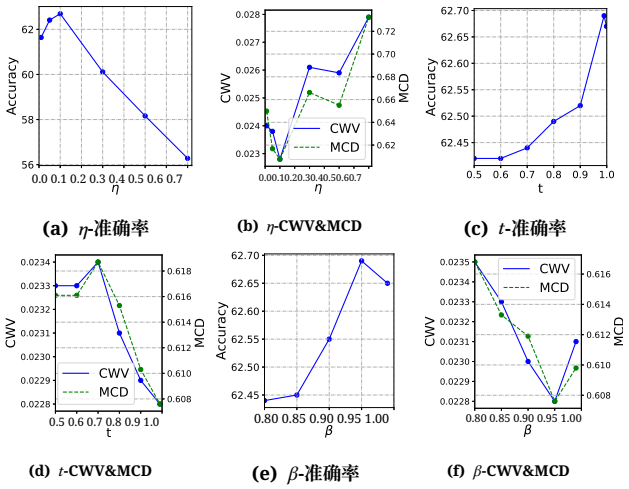
方法	平均准确率	平均 CWV	平均 MCD	最终准确率	最终CWV	最终MCD
Ciliate	62.69	0.0228	0.6076	56.60	0.0256	0.6888
不使用样本选择	60.65	0.0261	0.6413	54.92	0.0287	0.7146
不使用验证	62.42	0.0233	0.6161	55.92	0.0278	0.7122
无平衡蒸馏	60.88	0.0268	0.6662	53.93	0.0281	0.7139
纯 Dropout	61.60	0.0249	0.6214	55.73	0.0267	0.7009
纯普通方法	61.52	0.0248	0.6336	56.30	0.0268	0.7022
WA	63.36	0.0740	0.6492	56.17	0.0641	0.7365
WA 配合 Dropout	63.29	0.0751	0.6411	56.22	0.0638	0.7298
BiC	58.75	0.0637	0.6968	50.88	0.0532	0.7343
BiC 配合 Dropout	58.90	0.0644	0.7016	50.93	0.0540	0.7234

这表明，我们提出的数据集和训练优化技术有助于提升修复性能。

数据集精炼的影响：表 3 的前三行展示了对比结果。通过样本选择，平均准确率提高了 2.90%，公平性表现也有所提升，CWV 和 MCD 分别提高了 9.66% 和 3.97%。通过验证，Ciliate 的模型性能在准确率、CWV 和 MCD 方面分别提高了 0.43%、2.15% 和 1.38%。

训练优化的影响：表 3 的第一、第四、第五和第六行展示了不同训练优化的结果
方法。我们的平衡蒸馏分别将模型性能在准确率、CWV 和 MCD 方面提升了2.97%、14.93% 和 8.80%。我们的选择性训练方法在准确率、CWV和 MCD 方面分别超越常规训练 1.45%、4.03% 和3.97%，同时分别超越纯Dropout训练 1.31%、4.62% 和0.74%。这证实了Ciliate 中的选择性训练能够实现较高的准确率和公平性。表3的最后四行展示了将丢弃技术应用于 BiC 和 WA 的影响。丢弃技术使WA 的准确率、CWV 和 MCD 分别变化了 -0.11%、-1.49% 和 1.25%，同时使BiC 的准确率、CWV 和 MCD 分别变化了 0.26%、-1.10% 和 -0.69%。

分析： 从结果中，我们可以得到一些启示。首先，与其他方法相比，Ciliate取得了最高的效用和公平性表现，这证明了Ciliate各个阶段的有效性。其次，样本选择对性能的影响更大，而验证的影响最小，这为我们在§4.4中调整参数提供了指导。最后，对WA和BiC应用Dropout并未显著提升性能，这表明Dropout需要与数据集精炼相结合才能发挥作用。



(a) η-准确率 (b) η-CWV&MCD (c) t-准确率

(d) t-CWV&MCD (e) β-准确率 (f) β-CWV&MCD

图 11： η、 t和 β对模型平均性能的影响。**表 4：数据集精炼中使用的不同散度度量的比较。**

指标	平均准确率、平均CWV、平均MCD、最终准确率、最终CWV、最终MCD					
Jensen-Shannon散度	62.69	0.0228	0.6076	56.60	0.0256	0.6888
Kullback-Leibler散度	60.74	0.0239	0.6297	54.32	0.0262	0.7206
Hellinger距离	59.89	0.0252	0.6337	53.97	0.0282	0.7178

4.4 RQ3：可调配置的影响

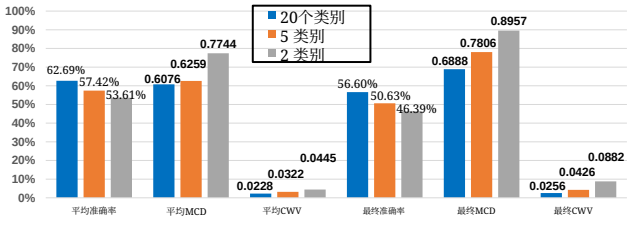


图 12：不同类别数量划分之间的比较。

实验设计： Ciliate 的实现需要预先确定超参数以及用于数据集精炼的度量指标。在本节中，我们研究这些配置如何影响修复性能。为此，1）我们针对验证超参数 t和 β分别在区间 $[0.5, 1.0]$ 和 $[0.9, 1.0]$内变化的情况进行对比实验；

2）我们进行了数据集精炼超参数 η 在区间 $[0.01, 0.7]$内变化的对比实验；3）除了Ciliate用于精炼数据集的Jensen-Shannon散度外，我们还使用其他指标评估性能，以衡量输出的相似性，包括库尔贝克-莱布勒散度和海灵格距离。4）我们在CIFAR-100上针对不同的增量类别数（2、5、20）进行了对比实验；我们遵循§4.2中的CIL设置，并在CIFAR-100 数据集上进行实验。我们比较这些方法的平均性能（不包括基础模型）以及最后一步获得的模型性能。

结果：我们分别考察超参数 η、t 和 β 的影响。图 11 展示了超参数设置对模型性能的影响，其中报告的是平均结果。此外，我们还研究了数据集精炼指标选择的影响。表 4 汇总了采用不同数据集精炼指标得到的结果。

超参数的影响 η：Ciliate 利用可配置的超参数 η 来控制数据集精炼的程度。η 表示数据集划分的阈值。随着其值的增加，更多样本会被归类为杂质样本，从而得到更小的精炼数据集。正如图11所示，随着 η 的增加，模型的性能先提升后下降。当 $\eta = 0.01$时，Ciliate 的表现最佳，因此我们选择该值作为默认设置。

超参数的影响 t 和 β：Ciliate 利用两个可配置的超参数，t和 β，来调整神经元覆盖率验证过程的严格程度。t 表示单个神经元的激活阈值，β 表示模型的神经元覆盖率阈值。如图 11 所示，随着 t 和 β 增加，模型的性能也会先提升后下降，并在 $t = 0.99$ 和 $\beta = 0.95$时达到最佳。因此，我们选择这些值作为默认设置。

数据集精炼的效果指标：从表4中，我们可以观察到，使用詹森-香农散度的模型在准确率、CWV和MCD方面分别比使用库尔贝克-莱布勒散度的模型高出2.75%、0.42%和2.05%，而分别比使用海林格距离的模型高出4.21%、5.56%和2.67%。这证实了在Ciliate中使用詹森-香农散度可以实现最佳的准确率和公平性。

增量类别数量的影响：为研究类别数量的影响，我们按照不同粒度划分数据集。我们将 CIFAR-100 划分为包含 20、5 和 2 个类别的批次，分别对应 5、20 和 50 个增量学习步骤。从图 12 中可以观察到，在 20个类别划分下的 Ciliate 相较于 5 个类别划分，在准确率、CWV 和 MCD 方面分别高出 8.41%、41.23% 和 7.46%；相较于 2 个类别划分，则分别高出 14.48%、95.18% 和 27.45%。这表明，增量学习所经历的步骤越多，模型性能的退化越严重。

分析：首先，从图11可以看出，随着超参数的增大，模型的性能先提升后下降，这意味着存在极值点。由于 η 是控制高质量数据集规模的超参数，我们可以说，高质量数据集的性能与其规模并不呈正相关。在我们看来，如果高质量数据集的规模过小，就难以涵盖所有重要特征；如果规模过大，则会引入更多噪声以及样本扰动，这使得模型难以提取真正重要的特征。同样地，如果 t和 b 的值过大，那么验证约束将过于严格，从而导致数据采样不足；否则，可能会导致特征表示不足。

最后，表 4 表明，模型的性能会直接受到散度度量选择的影响。Hellinger 模型的表现最差，我们认为这是因为它在计算 Hellinger 距离时使用了大量平方和平方根运算，这可能会导致数值问题。Jensen-Shannon 散度的性能优于 Kullback-Leibler 模型。我们认为其原因在于，Kullback-Leibler 散度的计算具有不对称性，即 $KL(P, Q)$ 不等于 $KL(Q, P)$，这使得模型难以准确衡量样本在训练中的重要性。

准确率-公平性权衡：模型准确率-公平性权衡 [39] 指的是，在给定任务上提高机器学习模型的准确率，可能会以其预测结果的公平性或公正性降低为代价。在许多情况下，为优化准确率而训练的模型会学会依赖数据中的歧视性特征，例如种族或性别，并作出有偏差的预测。相反，试图确保模型预测结果的公平性或公正性，可能会导致准确率降低，因为模型可能被迫忽略相关特征，或降低其权重。请注意，在不存在此类冲突的场景中，权衡可能并不存在 [24, 77, 83]。当某项任务的相关特征在不同人口群体之间均匀分布时，就可能出现这种情况，例如在某些情况下，特征分布并未偏向任何特定群体。在这些情况下，为优化给定任务上的准确率而训练的模型也可能是公平的，因为最相关的特征不具有歧视性。我们研究了 Ciliate 中是否存在这种权衡。η 的值，即数据集精炼程度，限制了公平性要求。在极端情况下，如果不施加公平性约束（$\eta = 0$），Ciliate 会将所有样本视为高质量样本，反之亦然。因此，我们通过分析准确率与 η之间的关系来研究准确率-公平性权衡。图 11 (a)&(b) 显示，随着 η 增长，准确率和公平性呈现相同的趋势（即先上升，随后因过拟合而下降）；实验使用了 CIFAR-100 数据集和 ResNet-32 模型。在 Flowers 和 Cars 上进行的实验显示了类似结果。请注意，权衡效应取决于不同的设置（例如数据集、模型），在其他设置下结果可能有所不同。

4.5 研究问题4：对增量学习的启示

基于上述研究和讨论，我们可以提出一些建议，以提升类别增量学习模型的性能。

增加采样数据集的规模。正如我们在 §3.2.1 中所解释的，保留更多旧样本至关重要，因为这可以显著减轻数据集偏差。我们的观察表明，对于 CIFAR-100 数据集，当旧样本数量增加到 5000（占训练样本总数的 10%）时，模型性能的提升已不明显。具体而言，当数量达到 5000 时，模型效用与最佳效用（存储所有旧样本）相差不超过 2%；相比之下，数量从 500 增至 5000 时，性能提升近 10%。这表明，增加采样数据集的规模有助于提升模型性能，但边际效应较小。

表 5：训练模型所需时间。

数据集	Ciliate	BiC	WA	iCaRL
CIFAR-100	53784s	52386s	60377s	55347s
花卉	12835秒	12390秒	13659秒	13031秒
Cars	8446秒	8403秒	8576s	8498s
ImageNet	215873s	214896s	220089s	217831秒

保持模型类别平衡。模型各类别之间的数据量存在显著差异，会影响其公平性表现（如图 3 所示）。因此，有必要在训练前对训练数据集进行预处理，以保持模型类别平衡。

促使模型学习重要特征。 特征在传统机器学习公平性研究中占据最重要的地位，对于图像分类系统而言，特征仍然发挥着重要作用。在训练阶段，模型必须持续提取数据的特征。如果模型在训练过程中遗漏了某些关键特征，其有效性可能会显著降低。这可能是由于难以从训练数据集中提取某些特征（如图 5 所示），或对这些特征学习不足（如图 7 和表 3 所示）所导致的。

从经验来看，采用这些策略有助于提升 CIL 模型的公平性表现和效用。我们认为，开发者适当地使用上述训练策略，可能会获得更好的增量学习流水线。

4.6 研究问题5：Ciliate 的效率

我们在 §4.2 所采用的相同设置下，使用 BiC、WA、iCaRL 和 Ciliate 测量增量学习训练的时间，以评估不同方法的效率。结果见表 5。第一列列出三个数据集，其余各列列出不同训练方法的时间开销。平均而言，Ciliate 比 BiC 多耗时 2.2%，比 WA 和 iCaRL 分别少耗时 6.7% 和 1.7%。总体而言，其耗时与其他方法没有显著差异，同时 Ciliate 实现了更好的修复性能。

5 对有效性的威胁

Ciliate 目前在 3 个数据集上进行了评估，这可能存在局限性。类似地，Ciliate中使用了可配置参数，尽管我们的实验表明，这些参数足以取得较高的修复效果，但当模型规模显著增大或缩小时，这一结论可能不再成立。为减轻这些威胁，所有原始和修复后的训练脚本、模型架构与训练配置详情、包括依赖项在内的实现，以及评估数据均可在 [1] 获取，以供复现。

6 相关工作与讨论

类别增量学习。 类别增量学习近年来成为一个活跃的研究主题。许多工作 [58, 82] 尝试在无法访问已见数据的情况下训练模型，但性能并不理想。主流策略主要可通过表征学习和分类器学习进行考察，它们基于数据记忆库有限的回放方法。

表征学习。以下三类通常用于对当代研究成果进行分类。基于正则化的方法 [6, 18, 41, 45, 85] 估计关键参数的变化，然后依次更新模型参数的后验分布。然而，其计算通常需要在很强的模型前提下进行近似。基于蒸馏的方法 [17, 20, 33, 66, 78, 89] 利用知识蒸馏 [32] 来维持表征。iCaRL [66]和 EE2L [17]在网络输出上计算蒸馏损失。UCIR [33] 不使用网络预测，而是使用归一化特征向量来计算蒸馏损失。PODNet [20]通过使用基于空间的蒸馏损失来限制模型的变化。基于结构的方法 [36, 81] 会固定与先前类别相关的已学习参数，并通过各种方式分配新参数，例如使用未使用的参数或额外的网络来学习新知识。

分类器学习。由于内存限制，类别不平衡问题是分类器学习方法面临的主要挑战。一些研究（如 LwF.MC [48] 和 RWalk [18], ）在单个训练阶段中共同训练特征提取器和分类器。

机器学习公平性。 随着自动化决策方法和系统的应用日益广泛，例如高等教育中的标准化考试 [19], 就业 [28, 65, 74], 以及再犯风险判定 [10, 11, 14, 57]，机器学习中的公平性问题引起了广泛关注。此外，政府（例如欧盟 [75]和美国 [55, 56]）、组织 [51], 以及媒体都呼吁提高公众责任意识，并增进社会对机器学习的理解。

为应对上述关切，机器学习模型的公平性测试成为一个重要的研究方向，其方法大多基于生成技术。THEMIS通过因果分析考量群体公平性，并采用随机测试生成来评估公平性 [7]。AEQUITAS专注于生成个体歧视实例 [73]。随后，ADF结合全局搜索和局部搜索，在梯度 [88]的引导下系统地搜索输入空间。符号生成（SG）整合符号执行和局部模型解释技术，以构造个体歧视实例 [4]。

发现公平性问题后，需要修复 ML 模型。为减轻数据集偏差，人们提出了预处理方法，包括纠正标签 [38, 87], 修订属性 [23,39]，生成无歧视数据 [67, 80]，以及获取公平数据表示 [12]。为减轻算法偏差，人们提出了许多处理过程中和后处理方法。具体而言，这些方法应用公平性 [21, 84], 提出考虑预测公平性的目标函数 [86], 设计新的训练框架 [3, 27, 80], 或直接更改有偏模型输出的预测标签 [29, 63]。Zheng 等人提出了 NeuronFair，这是一种深度神经网络公平性测试框架，可识别有偏神经元，以这些神经元为指导生成歧视性样本作为种子样本，并扰动种子样本以生成更多实例 [90]。基于线性回归的训练数据调试（LTDD）是另一种公平性测试工具。它侧重于检测哪些数据特征及其哪些部分存在偏差 [47]。Peng 等人提出了 xFAIR，这是一种基于模型的公平性修复方法，它利用数据特征之间的相关性来减轻偏差并解释其成因 [61, 62]。

讨论。 我们的方法 Ciliate 基于知识蒸馏，与 BiC 和 WA 有相似之处。然而，与这些方法不同，Ciliate 不会向模型引入额外的层或分类器，因为这可能会对模型的泛化能力产生负面影响。BiC 和 WA 旨在提高增量学习的准确率，而 Ciliate 则专注于提升公平性，同时避免准确率显著下降。目前，还没有其他增量学习方法对模型的公平性指标进行评估和约束，因此 Ciliate 是首个解决此问题的方法。

IL模型中的公平性问题。我们的目标是提高人们对增量学习系统中公平性问题的认识，并减轻其潜在的负面社会影响。与 BiC 和 WA 相比，Ciliate 在准确率表现相当的同时，显著提升了公平性表现。Ciliate 的计算成本高于 BiC，但低于WA；在未来的工作中，我们计划探索提高其效率的方法。总而言之，Ciliate 与现有方法的不同之处在于，它不增加额外层，并且系统地评估增量学习模型的公平性，使其成为表现更强、且在效率优化方面具有进一步潜力的方法。视觉变换器（ViT）最近成为一种新兴的视觉模型。受硬件条件限制，我们没有在 ViT上测试我们的方法，但我们认为数据集优化仍然适用，因为该方法不依赖于模型架构。将我们的方法应用于 ViT 将是未来一项有趣的工作。

7 结论

受软件调试的启发，我们提出并开发了 Ciliate，这是一种由数据集和训练优化驱动的自动化基于类别的增量学习模型公平性调试技术。它能够识别重要样本，并使用去偏差训练方法对这些样本进行模型训练。评估结果表明，Ciliate 构建的高质量数据集能够有效修复基于类别的增量学习中的模型公平性缺陷。

参考文献

[1] [n. d.]. 匿名代码库 - 匿名 GitHub。https://anonymous.4open.science/r/CILIATE-2C80 [2] Apple 机器学习研究 [n. d.]. 嘿 Siri：由设备端深度神经网络驱动的 Apple 个人助理语音唤醒功能。Apple 机器学习研究。https://machinelearning.apple.com/research/hey-siri [3] Tameem Adel、Isabel Valera、Zoubin Ghahramani 和 Adrian Weller。 [n. d.]. 单网络对抗公平性。33（n. d.），2412–2420。https://doi.org/10. 1609/aaai.v33i01.33012412 [4] Aniya Agarwal、Pranay Lohia、Seema Nagar、Kuntal Dey 和 Diptikalyan Saha。[n. d.]. 自动化测试生成：检测人工智能模型中的个体歧视。（n. d.）。arXiv: 1809 03260。Chirag Agarwal、Daniel D'souza 和 Sara Hooker。 [n. d.]. 使用梯度方差估计样本难度。https://doi.org/10.48550/arXiv.2008.11600 arXiv:2008.11600 [cs] [6] Rahaf Aljundi、Francesca Babiloni、Mohamed Elhoseiny、Marcus Rohrbach 和 Tinne Tuytelaars。 [n. d.]. 内存感知突触：学习不该遗忘的内容。载于《欧洲计算机视觉会议论文集》（ECCV）（2018）。139–154。[7] Rico Angell、Brittany Johnson、Yuriy Brun 和 Alexandra Meliou。 [n. d.]. THEMIS：自动测试软件中的歧视问题。载于《2018年第26届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议论文集》（ESEC/FSE 2018）（美国佛罗里达州莱克布埃纳维斯塔，2018 871 875 10 1145 3236024 3264590）。[8] ACM出版社，–。https://doi.org/ 。/。Fatma Başak Aydemir 和 Fabiano Dalpiaz。n. d.。伦理意识软件工程路线图。载于《软件公平性国际研讨会论文集》（美国纽约州纽约市，2018-05-29）（FairWare '18）。计算机协会，15 21 10 1145 3194770 3194778 –。https://doi.org/ 。/。[9] Yoshua Bengio、Jérôme Louradour、Ronan Collobert 和 Jason Weston。 [n. d.]. 课程学习。载于《第26届国际机器学习年度会议论文集》（2009）。41–48。[10] Richard Berk。 [n. d.]. 少年司法风险评估中的准确率与公平性。16, 1（n. d.），175–194。[11] Richard Berk、Hoda Heidari、Shahin Jabbari、Michael Kearns 和 Aaron Roth。[n. d.]. 刑事司法风险评估中的公平性：现状。50, 1（n. d.），3–44。[12] Alex Beutel、Jilin Chen、Zhe Zhao 和 Ed H. Chi。 [n. d.]. 对抗式学习公平表征时的数据决策与理论启示。（n. d.）。arXiv:1707.00075 [cs] http://arxiv.org/abs/1707.00075 [13] Mariusz Bojarski、Davide Del Testa、Daniel Dworakowski、Bernhard Firner、Beat Flepp、Prasoon Goyal、Lawrence D. Jackel、Mathew Monfort、Urs Muller 和 Jiakai Zhang。 [n. d.]. 自动驾驶汽车的端到端学习。（n. d.）。arXiv:1604.07316 [14] Tim Brennan 和 William L. Oliver。 [n. d.]。《犯罪学中机器学习技术的兴起：数据与研究问题复杂性的影响》。12（n. d.），551。[15] Yuriy Brun 和 Alexandra Meliou。 [n. d.]。《软件公平性》。载于《2018年第26届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议论文集》（美国纽约州纽约市，2018-10-26）（ESEC/FSE 2018）。计算机协会，754–759。https://doi.org/ 10 1145 3236024 3264838 。/ 。Mateusz Buda、Atsuto Maki 和 Maciej A. Mazurowski。 [n. d.]。《卷积神经网络中类别不平衡问题的系统性研究》。106（n. d.），249–259。[16] Francisco M. Castro、Manuel J. Marín-Jiménez、Nicolás Guil、Cordelia Schmid 和 Karteek Alahari。 [n. d.]。《端到端增量学习》。载于《欧洲计算机视觉会议论文集》（ECCV）（2018）。233–248。[17] Arslan Chaudhry、Puneet K. Dokania、Thalaiyasingam Ajanthan 和 Philip HS Torr。 [n. d.]。《用于增量学习的黎曼漫步：理解遗忘与顽固性》。载于《欧洲计算机视觉会议论文集》（ECCV）（2018）。532–547。[18] T. Anne Cleary。 [n. d.]。《测试偏差：学术能力倾向测验对综合院校黑人与白人学生的效度》。1966，2（n. d.），i–23。[19] Arthur Douillard、Matthieu Cord、Charles Ollion、Thomas Robert 和 Eduardo Valle。 [n. d.]。《Podnet：面向小任务增量学习的池化输出蒸馏》。载于《欧洲计算机视觉会议》（2020）。Springer，86–102。[20] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。 [n. d.]。《通过认知实现公平》。载于《第3届理论计算机科学创新会议论文集》（ITCS '12）（马萨诸塞州剑桥，2012 214 226 10 1145 2090236 2090255）。ACM出版社，–。https://doi.org/。/。[22] Dumitru Erhan、Yoshua Bengio、Aaron Courville 和 Pascal Vincent。 [n. d.]。《深度网络高层特征的可视化》。1341，3（n. d.），1。[23] Michael Feldman、Sorelle A. Friedler、John Moeller、Carlos Scheidegger 和 Suresh Venkatasubramanian。 [n. d.]。《认证并消除差别影响》。载于《第21届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集》（KDD '15）（澳大利亚新南威尔士州悉尼，2015）。ACM出版社，259 268 10 1145 2783258 2783311 –。https://doi.org/。/。[24] Benjamin Fish、Jeremy Kun 和 Ádám D. Lelkes。 [n. d.]。《一种平衡公平性与准确率的置信度方法》。https://doi.org/10.48550/arXiv。1601 05764 1601 05764 [ ]。arXiv：。cs Robert M. French。n. d.。《联结主义网络中的灾难性遗忘》。3，4 128 135（n. d.），–。[26] Sainyam Galhotra、Yuriy Brun 和 Alexandra Meliou。 [n. d.]。《公平性测试：检测软件中的歧视》。载于《2017年第11届软件工程基础联合会议论文集》（美国纽约州纽约市，2017-08-21）（ESEC/FSE 2017）。计算机协会，498–510。https://doi.org/ 10 1145 3106237 3106277 。/。Xuanqi Gao、Juan Zhai、Shiqing Ma、Chao Shen、Yufei Chen 和 Qian Wang。[n. d.]。《Fairneuron：通过选择性神经元上的对抗博弈提升深度神经网络公平性》。载于《2022 IEEE/ACM第44届软件工程国际会议》（ICSE）（2022）。921–933。https://doi.org/10.1145/3510003。3510087 [28] Robert M. Guion。 [n. d.]。《就业测试与歧视性招聘》。5，2（n. d.），20–37。[29] Moritz Hardt、Eric Price 和 Nati Srebro。 [n. d.]。《监督学习中的机会平等》。29（n. d.），3315–3323。[30] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。 [n. d.]。《用于图像识别的深度残差学习》。载于《IEEE计算机视觉与模式识别会议论文集》（2016）。770–778。[31] Ernst Hellinger。 [n. d.]。《无限多个变量的二次型理论的新论证》。1909，136（n. d.），210–271。[32] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。 [n. d.]。《蒸馏神经网络中的知识》。（n. d.）。arXiv: 1503 02531。Saihui Hou、Xinyu Pan、Chen Change Loy、Zilei Wang 和 Dahua Lin。 [n. d.]。《通过再平衡增量式学习统一分类器》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（2019）。831–839。[34] Yen-Chang Hsu、Yen-Cheng Liu、Anita Ramasamy 和 Zsolt Kira。 [n. d.]。《重新评估持续学习场景：分类方法及支持强基线的理由》。（n. d.）。arXiv: 1810 12488。Baotian Hu、Zhengdong Lu、Hang Li 和 Qingcai Chen。 [n. d.]。《用于匹配自然语言句子的卷积神经网络架构》。载于《神经信息处理系统进展》（2014）。2042–2050。[36] Ching-Yi Hung、Cheng-Hao Tu、Cheng-En Wu、Chien-Hung Chen、Yi-Ming Chan 和 Chu-Song Chen。 [n. d.]。《通过压缩、挑选与扩展实现持续学习中的遗忘消除》。32（n. d.）。[37] Nal Kalchbrenner、Edward Grefenstette 和 Phil Blunsom。 [n. d.]。《用于句子建模的卷积神经网络》。（n. d.）。arXiv: 1404 2188。Faisal Kamiran 和 Toon Calders。n. d.。《无歧视分类》。载于《2009年第2届计算机、控制与通信国际会议》（2009）。IEEE，1–6。

[39] Faisal Kamiran 和 Toon Calders。 [n. d.]。《无歧视分类的数据预处理技术》。 33，1（n. d.），1–33。https://doi.org/10. 1007/s10115-011-0463-8[40] Angelos Katharopoulos 和 Francois Fleuret。 [n. d.]。《并非所有样本都生而平等：采用重要性采样的深度学习》。载于《第35届国际机器学习会议论文集》（2018-07-03）。 PMLR，2525–2534。https://proceedings.mlr.press/v80/katharopoulos18a.html[41] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A. Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska、Demis Hassabis、Claudia Clopath、Dharshan Kumaran 和 Raia Hadsell。 [n. d.]。《克服神经网络中的灾难性遗忘》。（n. d.）。arXiv: 1612 00796 [ ] 1612 00796。[42] cs, stat http://arxiv.org/abs/。Jonathan Krause、Michael Stark、Jia Deng 和 Li Fei-Fei。n. d.]。《用于细粒度分类的三维物体表示》。载于《IEEE国际计算机视觉会议研讨会论文集》（2013）。554–561。[43] S. Kullback 和 R. A. Leibler。 [n. d.]。《论信息与充分性》。22，1（n. d.）， 79 86 10 1214 1177729694 –。https://doi.org/。/aoms/ M. [44] Kumar、Benjamin Packer 和 Daphne Koller。n. d.]。《潜变量模型的自步学习》。 23（n. d.）。[45] Sang-Woo Lee、Jin-Hwa Kim、Jaehyun Jun、Jung-Woo Ha 和 Byoung-Tak Zhang。[n. d.]。《通过增量矩匹配克服灾难性遗忘》。30（n. d.）。[46] Timothée Lesort、Vincenzo Lomonaco、Andrei Stoian、Davide Maltoni、David Filliat 和 Natalia Díaz-Rodríguez。 [n. d.]。《面向机器人的持续学习：定义、框架、学习策略、机遇与挑战》。58（n. d.），52–68。[47] Yanhui Li、Linghan Meng、Lin Chen、Li Yu、Di Wu、Yuming Zhou 和 Baowen Xu。[n. d.]。《用于保障机器学习软件公平性的训练数据调试》。载于《第44届国际软件工程会议论文集》（美国纽约州纽约市，2022-07-05）（ICSE '22）。计算机协会， 2215 2227 10 1145 3510003 3510091 –。https://doi.org/。/。[48] [ ] Zhizhong Li 和 Derek Hoiem。n. d.。《无遗忘学习》。40，12（n. d.）， 2935 2947 10 1109 2017 2773081 –。https://doi.org/。/[49] TPAMI.。Evan Z. Liu、Behzad Haghgoo、Annie S. Chen、Aditi Raghunathan、Pang Wei Koh、Shiori Sagawa、Percy Liang 和 Chelsea Finn。 [n. d.]。《只训练两次：在不使用群体信息训练的情况下提升群体鲁棒性》。载于《国际机器学习会议》（2021）。PMLR，6781–6792。[50] Christopher Manning 和 Hinrich Schutze。 [n. d.]。《统计自然语言处理基础》。MIT出版社。[51] Annette Markham 和 Elizabeth Buchanan。 [n. d.]。《伦理决策与互联网研究：2.0版。AoIR伦理工作委员会的建议》。（n. d.）。[52] Michael McCloskey 和 Neal J. Cohen。 [n. d.]。《联结主义网络中的灾难性干扰：序列学习问题》。载于《学习与动机心理学》。第24卷。 Elsevier，109–165。[53] Patrick McClure、Charles Y. Zheng、Jakub Kaczmarzyk、John Rogers-Lee、Satra Ghosh、Dylan Nielson、Peter A. Bandettini 和 Francisco Pereira。 [n. d.]。《分布式权重巩固：脑部分割案例研究》。31（n. d.）。[54] Maria-Elena Nilsback 和 Andrew Zisserman。 [n. d.]。《大量类别上的自动花卉分类》。载于《2008年第六届印度计算机视觉、图形学与图像处理会议》（2008）。 IEEE，722–729。[55] 总统行政办公室、Cecilia Munoz（国内政策委员会主任）、Megan Smith（美国首席技术官，科学技术政策办公室）以及 DJ Patil（数据政策副首席技术官、首席数据科学家，科学技术政策办公室）。 [n. d.]。《大数据：关于算法系统、机遇与公民权利的报告》。总统行政办公室。[56] 美国总统行政办公室和 John Podesta。 [n. d.]。《大数据：把握机遇，守护价值观》。白宫总统行政办公室。[57] Cathy O'neil。 [n. d.]。《数学毁灭性武器：大数据如何加剧不平等并威胁民主》。 Crown。[58] Oleksiy Ostapenko、Mihai Puscas、Tassilo Klein、Patrick Jahnichen 和 Moin Nabi。 [n. d.]。《学会记忆：一种由突触可塑性驱动的持续学习框架》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（2019）。11321–11329。[59] Daniele Palossi、Antonio Loquercio、Francesco Conti、Eric Flamand、Davide Scaramuzza 和 Luca Benini。 [n. d.]。《一款基于 DNN、功耗为64mW的自主纳米无人机视觉导航引擎》。6，5（n. d.），8357–8371。https: //doi.org/ 10 1109 2019 2917066 1805 01831 [ ]。/JIOT.。arXiv：。[60] [cs, eess Kexin Pei、Yinzhi Cao、Junfeng Yang 和 Suman Jana。n. d.]。《 DeepXplore：深度学习系统的自动化白盒测试》。载于《第26届操作系统原理研讨会论文集》（中国上海，2017-10-14）。ACM， 1 18 10 1145 3132747 3132785 –。https://doi.org/。/。[61] Kewen Peng、Joymallya Chakraborty 和 Tim Menzies。 [n. d.]。《FairMask：通过基于模型的受保护属性再平衡提升公平性》。https://doi.org/ 10.48550/arXiv.2110.01109 arXiv:2110.01109 [cs]

[62] Kewen Peng、Joymallya Chakraborty 和 Tim Menzies。 [n. d.]。xFAIR：通过基于模型的受保护属性再平衡提升公平性。（n. d.），arXiv: 2110 01109 [ ] 2110 01109 。cs http://arxiv.org/abs/ 。[63] Geoff Pleiss、Manish Raghavan、Felix Wu、Jon Kleinberg 和 Kilian Q Weinberger。[n. d.]。关于公平性与校准。（n. d.），10。[64] Novi Quadrianto、Viktoriia Sharmanska 和 Oliver Thomas。 [n. d.]。在数据域中发现公平表征。载于 IEEE/CVF计算机视觉与模式识别会议论文集（2019）。8227–8236。[65] Manish Raghavan、Solon Barocas、Jon Kleinberg 和 Karen Levy。 [n. d.]。缓解算法招聘中的偏差：评估相关主张与实践。载于2020年公平性、问责制与透明度会议论文集（ 2020）。469–481。[66] Sylvestre-Alvise Rebuffi、Alexander Kolesnikov、Georg Sperl 和 Christoph H. Lampert。 [n. d.]。Icarl：增量分类器与表征学习。载于IEEE 计算机视觉与模式识别会议论文集（2017）。2001–2010。[67] Prasanna Sattigeri、Samuel C. Hoffman、Vijil Chenthamarakshan 和 Kush R. Varshney。 [n. d.]。公平性生成对抗网络：使用生成对抗网络生成具有公平性属性的数据集。63，4/5（n. d.），3–1。[68] Or Sharir、Barak Peleg 和 Yoav Shoham。 [n. d.]。训练自然语言处理模型的成本：简要概述。（n. d.），arXiv:2004.08900[69] Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。 [n. d.]。Dropout：一种防止神经网络过拟合的简单方法。（n. d.），30。[70] Emma Strubell、Ananya Ganesh 和 Andrew McCallum。 [n. d.]。深度学习在自然语言处理中的能源与政策考量。（n. d.），arXiv: 1906 02243 。Qi Tian、[71] [ ] Kun Kuang、Kelu Jiang、Fei Wu 和 Yisen Wang。n. d.。对抗训练中类别级鲁棒性的分析与应用。（n. d.），arXiv:2105.14240[72] Antonio Torralba 和 Alexei A. Efros。 [n. d.]。客观看待数据集偏差。载于 CVPR 2011（2011）。IEEE，1521–1528。[73] Sakshi Udeshi、Pryanshu Arora 和 Sudipta Chattopadhyay。 [n. d.]。自动化定向公平性测试。载于第33届ACM/IEEE自动化软件工程国际会议论文集——ASE 2018（法国蒙彼利埃， 2018 98 108 10 1145 3238147 3238165 ）。ACM出版社，–。[74] [https://doi.org/。/. Elmira van den Broek、Anastasia Sergeeva 和 Marleen Huysman。n. d.]。招聘算法：实践中公平性的民族志。（n. d.）。[75] Paul Voigt 和 Axel Von dem Bussche。 [n. d.]。欧盟《通用数据保护条例》（GDPR）。10（n. d.），3152676。[76] Kellie Webster、Xuezhi Wang、Ian Tenney、Alex Beutel、Emily Pitler、Ellie Pavlick、Jilin Chen、Ed Chi 和 Slav Petrov。 [n. d.]。测量并减少预训练模型中的性别相关性。（n. d.），arXiv: 2010 06032 。Michael Wick 和[77] [ ] Jean-Baptiste Tristan。n. d.。解锁公平性：重新审视权衡。32（n. d.）。[78] Yue Wu、Yinpeng Chen、Lijuan Wang、Yuancheng Ye、Zicheng Liu、Yandong Guo 和 Yun Fu。 [n. d.]。大规模增量学习。https://doi.org/10.48550/arXiv. 1905 13260 1905 13260 [ ]。arXiv:。cs Xiaofei Xie、Simon See、Lei [79] Ma、Felix Juefei-Xu、Minhui Xue、Hongxu Chen、Yang Liu、Jianjun Zhao、Bo Li 和 Jianxiong Yin。 [n. d.]。DeepHunter：一种面向深度神经网络的覆盖率引导模糊测试框架。载于第28届ACM SIGSOFT软件测试与分析国际研讨会论文集——ISSTA 2019（中国北京，2019）。ACM出版社，146–157。 https://doi.org/10.1145/3293882.3330579[80] Depeng Xu、Shuhan Yuan、Lu Zhang 和 Xintao Wu。 [n. d.]。Fairgan：具备公平性意识的生成对抗网络。载于 2018年IEEE大数据国际会议（Big Data）（2018）。IEEE，570–575。[81] Shipeng Yan、Jiangwei Xie 和 Xuming He。 [n. d.]。DER：用于类别增量学习的动态可扩展表征。3014–3023。https://openac cess.thecvf.com/content/CVPR2021/html/Yan_DER_Dynamically_Expandable_ Representation_for_Class_Incremental_Learning_CVPR_2021_paper.html[82] Lu Yu、Bartlomiej Twardowski、Xialei Liu、Luis Herranz、Kai Wang、Yongmei Cheng、Shangling Jui 和 Joost van de Weijer。 [n. d.]。类别增量学习的语义漂移补偿。载于IEEE/CVF计算机视觉与模式识别会议论文集（2020）。6982–6991。[83] Muhammad Bilal Zafar、Isabel Valera、Manuel Gomez Rodriguez 和 Krishna P. Gummadi。 [n. d.]。超越差别待遇与差别影响的公平性：在没有差别不当对待的情况下学习分类。载于第26届万维网国际会议论文集（日内瓦共和国与州，CHE，2017-04-03）（WWW '17）。国际万维网会议指导委员会， 1171 1180 10 1145 3038912 3052660 –。[84] https://doi.org/。/. Muhammad Bilal Zafar、Isabel Valera、Manuel Gomez Rogriguez 和 Krishna P. Gummadi。 [n. d.]。公平性约束：实现公平分类的机制。载于人工智能与统计学（2017）。PMLR，962–970。[85] Friedemann Zenke、Ben Poole 和 Surya Ganguli。 [n. d.]。通过突触智能实现持续学习。载于国际机器学习会议（2017）。PMLR，3987–3995。[86] Brian Hu Zhang、Blake Lemoine 和 Margaret Mitchell。 [n. d.]。通过对抗学习缓解不良偏差。载于2018年AAAI/ACM人工智能、伦理与社会会议论文集（美国路易斯安那州新奥尔良，2018-12-27）。ACM，

335 340 10 1145 3278721 3278779 – 。https://doi.org/ 。 / 。Lu [87] Zhang、YongkaiWu和 Xintao Wu。未注明日期。[ ]。实现数据发布中的非歧视性。载于第*23*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集（加拿大新斯科舍省哈利法克斯，2017年-08-13）。ACM，1335–1344。https://doi.org/10.1145/3097983.3098167[88] Peixin Zhang、JingyiWang、Jun Sun、Guoliang Dong、Xinyu Wang、XingenWang、JinSong Dong和 Ting Dai。 [未注明日期。]。通过对抗性采样进行白盒公平性测试。载于*ACM/IEEE*第*42*届软件工程国际会议论文集（韩国首尔，2020-06-27）。ACM，949–960。 https://doi.org/ 10 1145 3377811 3380331 . / . Bowen Zhao、Xi Xiao、Guojun Gan[89]、Bin Zhang 和 Shu-Tao Xia。 [n. d.]。在类别增量学习中保持无歧视与公平性。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集(2020)。13208–13217.[90] Haibin Zheng、Zhiqing Chen、Tianyu Du、Xuhong Zhang、Yao Cheng、Shouling Ji、Jingyi Wang、Yue Yu 和 Jinyin Chen。 [n. d.]。NeuronFair：通过识别有偏差的神经元实现可解释的白盒公平性测试。arXiv:2112.13214 [cs] http://arxiv.org/abs/2112.13214