面向在线连续学习的特征融合引导的梯度重加权算法

邱奔流,王岚晓,邱荷茜,高翔宇,闫海涛,李宏亮
(电子科技大学信息与通信工程学院,四川成都 611731)

摘 要: 在线连续学习(Online Continual Learning,OCL)旨在从非平稳的数据流中以仅仅读取一次数据样本的方式学习知识,因此面临着学习不充分的问题. 为缓解这一问题,本文提出了一种特征融合的方法. 该方法提取每张图片的一系列增强后样本的特征作为锚点特征,并通过加权求和的操作来融合这些特征以获得融合特征. 融合权值由锚点特征和选定的作为枢纽特征的图片特征之间的相似性来决定. 优化这一融合特征的交叉熵损失能够促进学习进程,进而在当前新任务上取得更好的表现. 另外,我们提出了一致性损失来限制融合特征和枢纽特征之间的均方误差,以进一步提高模型在新任务上的表现. 最后,我们理论分析了交叉熵损失关于模型参数的梯度. 这一分析揭示了特征融合和梯度重加权之间的关系. 我们选择了在线连续学习的三个常用基准进行了大量的实验,包括 CIFAR-10、CIFAR-100 和 Tiny-ImageNet. 相比基准方法,本文方法的平均最终准确率在 CIFAR-10 上提高了至多 7.00%,在 CIFAR-100 上提高了至多 8.04%,在 Tiny-ImageNet 上提高了至多 6.33%. 实验结果表明了本文方法的有效性,并且其在线连续学习能力相比已有方法取得了实质性的提升.

关键词: 图像识别;连续学习;在线学习;类别增量学习;特征融合;梯度重加权
基金项目: 新一代人工智能国家科技重大专项(No.2021ZD0112001)
中图分类号: TP181;TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)11-3970-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250413

Gradient Re-Weighting Guided by Feature Fusion for Online Continual Learning

QIU Ben-liu, WANG Lan-xiao, QIU He-qian, GAO Xiang-yu, WEN Hai-tao, LI Hong-liang
(*School of Information and Communication Engineering*, *University of Electronic Science and Technology of China*, *Chengdu*, *Sichuan* 611731, *China*)

Abstract: Online continual learning (OCL) aims at learning a non-stationary data stream in a way of reading each data sample only once, and hence suffers from insufficient learning. To address this problem, we propose a feature fusion method in this work. Our method leverages augmented samples of an image for producing anchor features, and incorporates them to obtain a fused feature via a weighted summation operation. The weights are determined by the similarity between anchor features and a pre-designated pivotal feature of the image. Optimizing the cross-entropy loss of this fused feature can accelerate the learning process, resulting in better performance on the current task. Additionally, we propose a consistency loss that restricts the mean-square error between the fused feature and the pivotal feature, which can further improve the performance on the current task. Finally, we provide a theoretical analysis about the gradients of cross-entropy loss to model parameters. This analysis reveals the relationship between the feature fusion and the gradient re-weighting. Extensive experiments are conducted on three benchmarks under OCL settings, including CIFAR-10, CIFAR-100 and Tiny-ImageNet. Our method surpasses baselines at most 7.00%, 8.04%, 6.33% for average end accuracy on CIFAR-10, CIFAR-100 and Tiny-ImageNet, respectively. Experimental results demonstrate the proposed method is effective, and achieves substantial improvement over previous methods for online continual learning.

Key words: image recognition; continual learning; online learning; class incremental learning; feature fusion; gradient re-weighting
Foundation Item(s): National Science and Technology Major Project (No.2021ZD0112001)

收稿日期:2025-05-24;录用日期:2025-11-20;责任编辑:覃怀银

1 引言

连续学习(Continual Learning,CL)在自动驾驶[1~3]、具身智能[4~6]等领域具有广泛的应用前景.连续学习旨在赋予机器类似于人类的从逐步观察到的数据流中学习和记忆知识的能力.根据任务标签的可得性和不同任务间类别是否重叠,连续学习通常可被分为任务增量型,域增量型和类别增量型[7].当前,大多数连续学习方法在离线的设置下训练模型,也就是模型可在完整的训练数据上学习多轮,也因此更容易遗忘已学习的知识.然而,在线设置在真实应用场景中更为普遍和实用.在线设置下,每个样本仅仅被训练一次,因此要求机器能够有效地处理数据流.

现有工作大多数从重放旧任务样本的角度来实现在线连续学习[8~12](Online Continual Learning,OCL).这一重放的策略选择并存储训练样本的一个子集,在学习新任务时再利用这些存储的数据来训练模型.除了直接存储样本,也能存储模型的中间表征和原型[13~15].一些工作使用知识蒸馏[16]来缓解不平衡问题[17~20],或者最大化互信息[21].结合了对比学习的数据增强方法也被用来学习表征性更强的特征[14,22].新的增强技术[21,23]也被针对性地提出以有效地利用样本中的信息.为增强每个样本的信息量,文献[24]使用了数据集压缩技术[25]来减少存储样本的数量.然而,这些现有的在线连续学习方法大多数专注于保留已学习知识,而忽视了对新知识的有效学习.

在线连续学习中,有效地从数据集中学习新知识有利于提升模型整体的表现,因为训练数据仅仅能被使用一次,也就是在一个训练批次中被使用.每个训练样本的一次性训练难以使得模型参数更新至最优状态.如何有效地从数据流中学习依旧是在线连续学习中富有挑战又充满前景的研究方向.本文提出了一种特征融合策略,以促进新任务样本的学习过程.图1展示了本文方法相比现有连续学习方法展现出更强的有效学习新任务的能力.该方法本质上为相应特征的梯度分配了不同的权重.具体来说,我们首先通过数据增强获取了每个图片样本的一组不同视角的样本.然后,通过一个可学习的特征提取器提取每个增强样本的特征.这些特征被视为每个图片样本的锚点特征.我们从每组图片样本的特征中预先选定一个特征作为该组样本的枢纽特征,并希望余下的特征收敛至这一枢纽特征.接着,我们根据锚点特征和枢纽特征之间的相似度来加权求和这些锚点特征以获取融合特征.这一融合特征吸收了更多对应图片的有关信息,并促进了图片特征的收敛.进一步,我们提出了一种一致性损失.该损失限制了融合特征和指定的枢纽特征之间的均方差误差,以提升学习效率.最后,我们提供了一个关于交叉熵损失对模型参数的更新梯度的理论分析.该分析指出,融合特征将关于分类器参数的梯度更多地推向枢纽特征,因此促进了锚点特征收敛于枢纽特征.本文在三个基准数据集CIFAR-10、CIFAR-100和Tiny-ImageNet上,在多种记忆缓存器容量下进行了实验.实验结果显示本文方法相比基准方法,平均最终准确率在CIFAR-10上提高至多7%,在CIFAR-100上提高至多8.04%,在Tiny-ImageNet上提高至多6.33%,表明本文方法有效地提升了模型的在线连续学习能力.

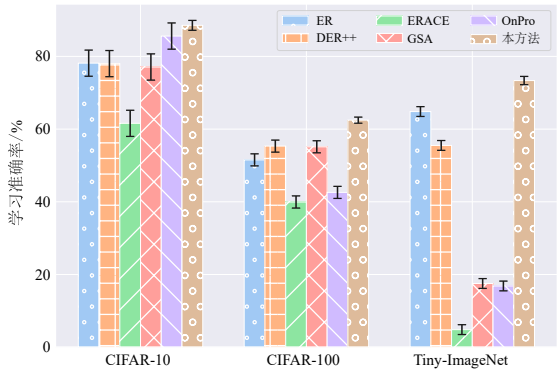


图1　本文方法和现有在线连续方法有效学习新任务能力对比

本文的主要贡献如下:

(1)提出了一种特征融合方法和一致性损失,将从增强图片集中提取的锚点特征集自适应吸纳进融合特征,并进一步对融合特征和枢纽特征之间的差异进行强约束,以有效地提升模型学习新任务的效率;

(2)提供了关于梯度的理论分析,指出特征融合方法等价于对锚点特征引起的交叉熵损失关于分类器参数的梯度进行重加权操作;

(3)根据本文方法在三个在线连续学习基准数据集上的实验,分析了模型的整体分类准确率、新任务的分类准确率、小存储容量适应性、重放方法适应性、超参数稳定性、训练时间长短、特征空间和融合权重可视化等,以研究本文方法的效度.

2 相关研究

2.1 连续学习

连续学习[26]旨在赋予模型从任务序列中逐渐学习知识的能力.主流的连续学习方法可以分为三类:重放、正则和动态结构.重放策略[9,10,18]在记忆缓存器或者生成模型中存储部分先前任务的样本,在学习当前任务时重新使用这些样本训练模型.正则策略[17,20,22]限制重要模型参数或者模型中间输出的变化量,以此来保留先前的知识.动态结构策略为每个任务拓展或者隔离出一部分网络参数.这些策略中,重放策略凭借其简洁性和有效性而得到广泛的应用.A-GEM[9]限制

了梯度方向使得其不会增加存储样本的平均损失. GDumb[10]采用贪婪策略存储样本,并保证不同类别的样本数量平衡. IL2A[23]使用了类别增强和语义增强,来学习表征和分类器. Co^{2}L[22]使用对比损失来学习表征,并使用自监督蒸馏来保持旧任务表征.

2.2 在线连续学习

在线连续学习是真实应用场景中更具有实际性的设置,该设置要求每个样本仅被学习一次. 这一在线设置满足了隐私保护和实时处理的要求. 大多数连续学习方法是为多阶段的离线学习场景而设计的,因此在处理在线连续学习时表现不佳. 现有在线连续学习方法主要采用重放策略,并取得了显著的结果. ER[8]指出存储部分先前任务的样本,并加入到当前任务的训练中可以带来好的表现. GSS[27]通过增大损失关于存储样本的梯度方向的方差,保持了存储样本的多样性. MIR[28]从记忆缓存中获取干涉最严重的样本,这些样本在估计的模型参数更新下更易导致损失增加. 在ER的基础上,DER++[13]存储了额外的样本logits,通过在训练损失中添加新的关于logits的正则项来约束模型更新. ASER[11]利用对抗Shapley值来给样本进行重要性打分. SCR[14]使用监督式对比损失来学习高质量特征. OCM[21]学习更为全面的特征,并通过最大化互信息量来保留先前学会的知识. ERACE[12]通过分离当前流数据的损失和存储数据的损失,改进了传统的交叉熵损失. GSA[29]识别了模型关于logits梯度的不平衡性,并提出了一种基于梯度的自适应损失来缓解不平衡问题. OnPro[15]指出在线连续学习中存在捷径学习(shortcut learning)的问题,并提出在线原型均衡和自适应原型反馈来学习更具判别性的特征. SSD[24]将来自训练数据流的知识总结成信息量更高的样本,并存储下来在后续任务中重放. 最近,研究者也探索了多目标优化技术[30]和等角紧致框架[31,32]于在线连续学习中的应用.

3 研究方法

3.1 基础知识

3.1.1 问题形成

本文中在线连续学习采取类别增量的设置. 训练数据$\mathcal{D}=\{\mathcal{D}_t\}_{t=1}^{T}$包含的任务序列共$T$个任务. 每个任务拥有数据$\mathcal{D}_t=\{(\boldsymbol{x}_i^t,y_i^t)\}(i=1,2,\cdots,N_t)$,其中$\boldsymbol{x}_i^t$代表输入图片,$y_i^t$代表相应的类别标签,$N_t$是任务$t$的样本数量. 我们使用$\mathcal{C}_t$代表任务$t$拥有的类别集合,$y_i^t\in\mathcal{C}_t$,$\mathcal{C}=\bigcup_{t=1}^{T}\mathcal{C}_t$. 由于类别增量设置下,任务间没有重合的类别,因此有$\mathcal{C}_i\cap\mathcal{C}_j=\varnothing,i\neq j$.

我们使用的分类模型$F=h\circ f$由两部分组成:一个带有参数$\boldsymbol{\Theta}$的特征提取器$h(\boldsymbol{x};\boldsymbol{\Theta})$和一个带有参数矩阵$\boldsymbol{W}=\left[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{|\mathcal{C}|}\right]$的分类器$f(\boldsymbol{z};\boldsymbol{W})$. 基于重放的方法通常使用记忆缓存器$\mathcal{M}$来存储少量的已见类别样本. 在当前任务$t$的训练过程中,从记忆缓存器$\mathcal{M}$中采样小批量的数据$X^{\text{buf}}$,从输入数据流$\mathcal{D}_t$中获取另一部分小批量的数据$X^{\text{new}}$. 这样,模型在数据$X^{\text{new}}\bigcup X^{\text{buf}}$上联合训练. 注意,这里每个样本仅训练一次,也就是数据流仅被训练一个阶段(epoch),以满足在线设置的要求. 模型产生的logits表示为$\{o_{c_i}\}(i=1,2,\cdots,|C|)$,这些logits进一步用来计算训练过程中的损失,或者在推断阶段用来预测类别标签.

3.1.2 梯度分析

基于logits值,样本图片$\boldsymbol{x}$属于类别c的概率p_c可以表示为

$$p_c=\frac{\mathrm{e}^{o_c}}{\sum_{j\in\mathcal{C}}\mathrm{e}^{o_j}},\ o_j=f\left(h(\boldsymbol{x};\boldsymbol{\Theta});\boldsymbol{W}\right)[j] \tag{1}$$

在任务t的训练阶段,模型仅仅能够获取新数据流中的数据$X^{\text{new}}\subseteq\mathcal{D}_t$和记忆缓存器中的数据$X^{\text{buf}}\subseteq\mathcal{M}$. 优化目标是最小化交叉熵损失,该损失可计算如下:

$$\mathcal{L}_{\text{ce}}=\mathbb{E}_{(\boldsymbol{x},y)\in X^{\text{new}}\cup X^{\text{buf}}}\left[-\log\left(\frac{\mathrm{e}^{o_y}}{\sum_{j\in\mathcal{C}}\mathrm{e}^{o_j}}\right)\right] \tag{2}$$

对于样本$(\boldsymbol{x},y)$,其损失$\mathcal{L}_{\text{ce}}$对分类器$f(\cdot)$的参数矩阵$\boldsymbol{W}$的梯度由下式给定:

$$\frac{\partial\mathcal{L}_{\text{ce}}}{\partial\boldsymbol{W}}=\begin{cases}h(\boldsymbol{x};\boldsymbol{\Theta})(p_i-1), & i=y\\ h(\boldsymbol{x};\boldsymbol{\Theta})(p_i), & i\neq y\end{cases} \tag{3}$$

对应类别i的分类器参数向量$\boldsymbol{w}_i$通过下述规则更新:

$$\boldsymbol{w}_i=\boldsymbol{w}_i-\eta\frac{\partial\mathcal{L}_{\text{ce}}}{\partial\boldsymbol{W}} \tag{4}$$

从式(3)中可知,对于属于类别y的样本$\boldsymbol{x}$,分类器向量$\boldsymbol{w}_y$的梯度为负值,而分类器向量$\boldsymbol{w}_i(i\neq y)$梯度为正值,这是因为激活层往往采用ReLU函数,并且有$0<p_i<1$. 根据式(4),考虑到学习率η为正数,负梯度会使得分类器向量$\boldsymbol{w}_y$偏向特征向量$h(\boldsymbol{x};\boldsymbol{\Theta})$. 相应地,正梯度值将分类器向量$\boldsymbol{w}_i(i\neq y)$拉离特征向量$h(\boldsymbol{x};\boldsymbol{\Theta})$. 基于反向传播的链式法则,损失关于分类器向量的这些梯度将传播至更浅的网络层,并更新特征提取器的参数,最终使得类别特征更靠近该类对应的分类器向量.

3.2 特征融合引导的梯度重加权

3.2.1 本文方法设计

根据上一节的分析可知,参数的更新方向由图片$\boldsymbol{x}$的特征$h(\boldsymbol{x};\boldsymbol{\Theta})$决定,因此特征的品质将很大程度上影响模型更新. 低品质的特征向量将在不合适的方向上

更新模型参数. 相较而言，高品质的特征向量将沿合适的方向促进参数更新，从而使得模型收敛更快、表现更优. 因此，一个自然的疑问是：我们能怎样提升高品质特征的正面影响，并弱化低品质特征的负面影响？

本文方法的整体框架如图2所示. 我们提出使用一个锚点特征的集合来减少低品质特征对相应分类器向量的负面影响，并增大高品质特征带来的正面影响. 我们使用数据增强来为每张输入图片$\boldsymbol{x}$产生少量锚点图片. 第i个数据增强操作表示为$\mathcal{A}_i(\cdot)$，并将第i张锚点图片表示为$\hat{\boldsymbol{x}}_i=\mathcal{A}_i(\boldsymbol{x})$. 然后，将这些锚点图片输入特征提取器$h(\boldsymbol{x};\boldsymbol{\Theta})$，提取的特征相应地表示为

$$\hat{z}_i=h\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\Theta}\right)=h\left(\mathcal{A}_i(\boldsymbol{x});\boldsymbol{\Theta}\right) \tag{5}$$

这些锚点特征不仅仅包含了能促进相应分类器向量收敛的高质量特征，也包含了减慢学习过程的低质量特征. 在特征空间中，通过对这些特征使用非负权值加权求和，我们能获得一个关于锚点特征的凸包. 具体地，我们表示所有锚点特征形成的矩阵为$\hat{\boldsymbol{Z}}=\left[\hat{z}_1,\hat{z}_2,\cdots,\hat{z}_m\right]$，其中$m$为锚点特征的数量. 权值向量相应地表示为$\boldsymbol{\alpha}$，锚点特征形成的凸包也就可表示为

$$\mathrm{CH}\left(\hat{\boldsymbol{Z}}\right)=\left\{\hat{\boldsymbol{Z}}\boldsymbol{\alpha}\middle|\forall\boldsymbol{\alpha},\boldsymbol{\alpha}\succ 0,\mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}=1\right\} \tag{6}$$

其中，$\mathbf{1}$为全1列向量. 文献[33]发现分类器向量在训练充分的理想情况下，将形成等角紧致框架结构，每类特征将收敛至特征空间中的一个点. 因此，我们定义特征空间中特征与类别特征中心之间的距离为特征品质的度量，也即更小的距离意味着更高的品质，反之亦然. 基于这一定义，我们可知对于权重$\boldsymbol{\alpha}$的多个取值，这些加权平均特征的品质将优于最低品质，如下式所示：

$$\mathcal{Q}\left(\hat{z},z^*\right)\geqslant\min_i\mathcal{Q}\left(\hat{z}_i,z^*\right),\forall\hat{z}\in\mathrm{CH}\left(\hat{\boldsymbol{Z}}\right) \tag{7}$$

其中，$\mathcal{Q}(\cdot,\cdot)$为一种品质度量；z^*为充分训练后获得的理想图片特征，也即类中心. 如果一个融合特征用来计算优化损失，并更新相关的分类器向量，它至少能缓解低品质特征导致的负面影响. 恰当的权重向量$\boldsymbol{\alpha}$的取值能够进一步使得分类器向量的更新更多地受益于高品质特征.

考虑到在线连续学习中非充分训练的情况下不可直接获取到理想图片特征，我们可指定输入图片的特征为理想图片特征的临时代理（见附录A）. 这一代理被称为枢纽特征，我们期望剩余的特征收敛至这一枢纽特征. 我们使用枢纽特征和余下增强图片的特征之间的相似性来计算权重向量$\boldsymbol{\alpha}$的取值. 具体地，权重向量$\boldsymbol{\alpha}$计算如下：

$$\boldsymbol{\alpha}(\boldsymbol{x})=\left[\alpha_1,\alpha_2,\cdots,\alpha_m\right]^{\mathrm{T}},\alpha_i=\frac{\mathrm{e}^{s\left(\hat{z}_i,z\right)}}{\sum_{j=1}^{m}\mathrm{e}^{s\left(\hat{z}_j,z\right)}} \tag{8}$$

其中，$s(\cdot,\cdot)$代表度量两向量相似性的点积操作；z为预先指定的枢纽特征.

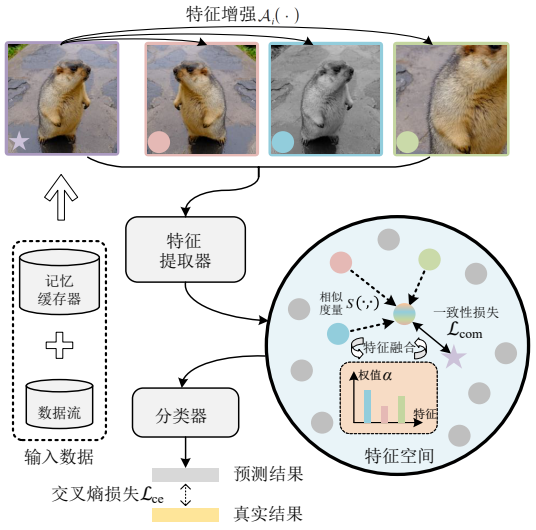


图2　本文方法的整体框架图

获取权值向量$\boldsymbol{\alpha}$之后，我们能进一步产生关于$\boldsymbol{x}$的融合特征$\hat{z}$，如下式所示：

$$\hat{z}=\hat{\boldsymbol{Z}}(\boldsymbol{x})\boldsymbol{\alpha}(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i\hat{z}_i\in\mathrm{CH}\left(\hat{\boldsymbol{Z}}\right) \tag{9}$$

其中，$\hat{\boldsymbol{Z}}(\boldsymbol{x})$是图片$\boldsymbol{x}$产生的锚点特征形成的矩阵. 这一融合特征接着送入分类器$f(\cdot;\boldsymbol{W})$来预测输入图片$\boldsymbol{x}$对应的类别y. 在训练过程中，这一融合特征对应的交叉熵损失为

$$\mathcal{L}_{\mathrm{ce}}\left(\hat{z}\right)=-\log\left(\frac{\mathrm{e}^{o_y}}{\sum_{j\in\mathcal{C}}\mathrm{e}^{o_j}}\right) \tag{10}$$

其中，$\left[o_{c_1},o_{c_2},\cdots,o_{c_{|\mathcal{C}|}}\right]$是分类器$f(\hat{z};\boldsymbol{W})$预测的logits值.

在上述特征加权融合方法的基础上，我们另外提出了一种一致性损失来促进特征的收敛，以提升模型有效学习新任务的能力. 这个一致性损失旨在限制融合特征和枢纽特征之间的距离. 我们选取具有直接强约束的均方误差函数. 该一致性损失定义如下：

$$\mathcal{L}_{\mathrm{con}}=\mathbb{E}_{(\boldsymbol{x},y)\sim X^{\mathrm{new}}\cup X^{\mathrm{buf}}}\|\hat{z}-h(\boldsymbol{x};\boldsymbol{\Theta})\|^2 \tag{11}$$

本文方法最终的完整损失可计算如下：

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{ce}}+\lambda\mathcal{L}_{\mathrm{con}} \tag{12}$$

其中，λ为控制一致性损失强度的超参数. 本文所提方法的完整流程如算法1所述.

3.2.2　关于梯度的理论分析

在这一部分，我们将讨论融合特征的梯度计算方式，并阐述特征融合与梯度重加权之间的关系. 根据3.1节中的讨论，属于类别y的融合特征$\hat{z}$的交叉熵损失关于分类器参数$\boldsymbol{W}$的梯度可以计算如下：

算法1 带有一致性损失的特征融合引导的梯度重加权算法

输入:数据流 $\mathcal{D}=\left\{\mathcal{D}_t\right\}_{t=1}^{T}$,特征提取器参数 $\boldsymbol{\Theta}$ 和分类器参数 $\boldsymbol{W}$,平衡因子 λ,学习率 η

输出:特征提取器参数 $\boldsymbol{\Theta}$ 和分类器参数 $\boldsymbol{W}$

1. 初始化记忆缓存器 $\mathcal{M}\leftarrow\varnothing$
2. FOR $t\in\{1,2,\cdots,T\}$ DO
3. FOR $X^{\text{new}}\sim\mathcal{D}_t$ DO
4. 采样训练样本 $(\boldsymbol{x},y)\sim X^{\text{new}}\cup X^{\text{buf}}$,其中 $X^{\text{buf}}\sim\mathcal{M}$
5. 获取锚点样本 $\hat{\boldsymbol{x}}_i\leftarrow\mathcal{A}_i(\boldsymbol{x}),\forall i=1,2,\cdots,m$,
6. 使用特征提取器提取枢纽特征 $\boldsymbol{z}\leftarrow h(\boldsymbol{x};\boldsymbol{\Theta})$ 和锚点特征 $\hat{\boldsymbol{z}}_i\leftarrow h(\hat{\boldsymbol{x}}_i;\boldsymbol{\Theta})$
7. 计算特征权值:$\alpha_i\leftarrow\dfrac{\mathrm{e}^{s(\hat{\boldsymbol{z}}_i,\boldsymbol{z})}}{\sum_{j=1}^{m}\mathrm{e}^{s(\hat{\boldsymbol{z}}_j,\boldsymbol{z})}},\forall i=1,2,\cdots,m$
8. 计算融合特征:$\hat{\boldsymbol{z}}\leftarrow\sum_{i=1}^{m}\alpha_i\hat{\boldsymbol{z}}_i$
9. 计算一致性损失:$\mathcal{L}_{\text{con}}\leftarrow\|\hat{\boldsymbol{z}}-\boldsymbol{z}\|^2$
10. 根据反向传播得到的梯度更新特征提取器和分类器参数:$(\boldsymbol{\Theta},\boldsymbol{W})\leftarrow(\boldsymbol{\Theta},\boldsymbol{W})-\eta\cdot\nabla_{(\boldsymbol{\Theta},\boldsymbol{W})}\left[\mathcal{L}_{\text{ce}}+\lambda\mathcal{L}_{\text{con}}\right]$
11. 更新记忆缓存器:$\mathcal{M}\leftarrow\text{Reservoir}\left(\mathcal{M},(\boldsymbol{x},y)\right)$
12. END FOR
13. END FOR
14. RETURN 特征提取器参数 $\boldsymbol{\Theta}$ 和分类器参数 $\boldsymbol{W}$

$$\frac{\partial\mathcal{L}_{\text{ce}}(\hat{\boldsymbol{z}})}{\partial\boldsymbol{W}}=\begin{cases}(p_i-1)\hat{\boldsymbol{z}}, & i=y\\ p_i\hat{\boldsymbol{z}}, & i\neq y\end{cases}\tag{13}$$

$$=\begin{cases}\sum_{j=1}^{m}\alpha_j(p_i-1)\hat{\boldsymbol{z}}_j, & i=y\\ \sum_{j=1}^{m}\alpha_j p_i\hat{\boldsymbol{z}}_j, & i\neq y\end{cases}\tag{14}$$

其中,p_i 表示将特征 $\hat{\boldsymbol{z}}$ 分类为类别 i 的概率.

对于任意单一特征 $\hat{\boldsymbol{z}}_j$,其对任意分类器向量 $\boldsymbol{w}_i$ 造成的梯度为

$$\frac{\partial\mathcal{L}_{\text{ce}}(\hat{\boldsymbol{z}}_j)}{\partial\boldsymbol{W}}=\begin{cases}(p_i'-1)\hat{\boldsymbol{z}}_j, & i=y\\ p_i'\hat{\boldsymbol{z}}_j, & i\neq y\end{cases}\tag{15}$$

其中,p_i' 代表 $\hat{\boldsymbol{z}}_j$ 分类为类别 i 的概率. 通过比较式(14)和式(15),我们能观察到融合特征梯度由各个锚点特征各自的梯度重加权得到,相应的权重具体表示如下:

$$\beta_j=\begin{cases}\alpha_j\dfrac{p_i-1}{p_i'-1}, & i=y\\ \alpha_j\dfrac{p_i}{p_i'}, & i\neq y\end{cases}\tag{16}$$

因此,融合特征 $\hat{\boldsymbol{z}}$ 的梯度可以重新表达如下:

$$\frac{\partial\mathcal{L}_{\text{ce}}(\hat{\boldsymbol{z}})}{\partial\boldsymbol{W}}=\sum_{j=1}^{m}\beta_j\frac{\partial\mathcal{L}_{\text{ce}}(\hat{\boldsymbol{z}}_j)}{\partial\boldsymbol{W}}\tag{17}$$

其中,我们可以直观地看到融合特征对梯度的影响等价于重加权它对应的各个锚点特征的梯度. 我们下面将给出这一重加权技术对学习过程影响的更为详细的分析. 对于一个靠近枢纽特征的高品质特征 $\hat{\boldsymbol{z}}_j$ 来说,其相比融合特征有更高的概率被正确分类,即 $p_i<p_i'$,可进一步得:

$$\frac{p_i}{p_i'}<1,\quad\frac{p_i-1}{p_i'-1}>1\tag{18}$$

这意味着该特征对分类器向量 $\boldsymbol{w}_i(i=y)$ 有更大的影响,以将其推向更靠近类别中心的位置,并且这一影响将被相对更大的 α_j 取值所放大. 同时,尽管这一特征对余下特征分类器向量 $\boldsymbol{w}_i(i\neq y)$ 的影响被弱化,但相对更大的 α_j 取值补偿了这一负作用. 另一方面,对于远离枢纽特征的低品质特征 $\hat{\boldsymbol{z}}_j$ 来说,其被正确预测的概率相比融合特征更小,也即 $p_i>p_i'$,进一步有:

$$\frac{p_i}{p_i'}>1,\quad\frac{p_i-1}{p_i'-1}<1\tag{19}$$

这表明该特征对更新分类器向量 $\boldsymbol{w}_i(i=y)$ 有更小的影响,并且这一影响进一步被更小取值的 α_j 给弱化. 相应地,该特征对分类器向量 $\boldsymbol{w}_i(i\neq y)$ 的负面影响也被更小的 α_j 取值所缓解. 总的来说,本文提出的特征融合方法提升了高品质样本的正面影响,同时缓解了低品质特征的负面影响,因而促进了在线连续学习场景下模型的学习过程.

4 实验及结果分析

4.1 实验设置

4.1.1 评估基准

本文方法的有效性在图片分类基准 CIFAR-10[34]、CIFAR-100[34] 和 Tiny-ImageNet[35] 上进行评估. CIFAR-10 包含了 10 类共 60 000 张训练样本,其中 50 000 张用于训练,10 000 张用于测试. CIFAR-100 包含了 100 个类别,每类 500 张样本用于训练,100 张样本用于测试. Tiny-ImageNet 共有 200 个类别,每类包含 500 张训练样本,50 张测试样本. 参照现有工作[29],我们将 CIFAR-10 划分为 5 个类别不相交的任务,每个任务包含 2 个类别;将 CIFAR-100 分成 10 个类别不相交的任务,每个任务包含 10 个类别;将 Tiny-ImageNet 分成 100 个类别不相交的任务,每任务包含 2 个类别. “在线”设置体现为模型训练过程中仅遍历数据集一遍,即每个数据仅在一个批数据中出现.

4.1.2 实现细节

考虑到不同方法间比较的公平性,我们参照现有工作[29],在三个数据集上均使用完整的 ResNet18[36] 网络作为特征提取器 h,并使用一个线性层作为分类器 f. 本文方法使用 AdamW 优化器从头开始训练模型. 学习

率设置为0.000 5,权值衰减值固定为0.000 1. 新数据流的批大小设为10,重放数据的批大小固定为64. 产生锚点样本的数据增强方式包括随机裁剪、水平翻转和随机增强. 实验主要基于NVIDIA TITAN V GPU.

4.1.3 比较方法

本文主要比较ER[8]、DER++[13]、ERACE[12]、GSA[29]、OnPro[15]、和SSD[24]. 这些方法中,ER和DER++既能处理在线连续学习,又能处理离线连续学习. 而ERACE、GSA、OnPro和SSD是为在线连续学习特殊设计的. 本文也列出了一些研究[21,29]报告的性能结果作为参考,包含的方法有GSS、MIR、ASER、GDumb、SCR、IL2A、Co^{2}L和OCM. 没有特别指出的话,本文方法以ER为基础,并采用它的记忆重放策略.

4.1.4 评估指标

平均最终准确率(average end accuracy)度量了模型的整体表现. 学习准确率(learning accuracy)评估了模型有效学习新任务的能力,两者分别计算如下:

$$\text{EndAcc}=\frac{1}{T}\sum_{j=1}^{T}a_j^T,\ \text{LAcc}=\frac{1}{T}\sum_{j=1}^{T}a_j^j \tag{20}$$

其中,a_j^i表示在任务i的训练集上完成训练后,在任务j的测试集上所得的准确率. 我们报告了10个不同随机种子下运行所得结果的均值和标准差.

4.2 主要结果

在线类别增量学习的设置下,在不同大小的存储缓存器尺寸下,本文方法和一些相关工作做了比较. 表1和表2报告了在CIFAR-10、CIFAR-100和Tiny-ImageNet数据集上的实验结果.

表1展示了多个方法的平均最终准确率. 从中可知,本文方法在大多数设置下取得了最高的平均最终准确率. 在CIFAR-100数据集上,在1 000、2 000、5 000的记忆缓存器尺寸下,分别超过第二优的方法0.42%、2.73%和5.94%. 在Tiny-ImageNet数据集上,在记忆缓存器尺寸为4 000和10 000的情况下,本文方法超过第二优的方法OnPro分别1.44%和4.66%的准确率. 尽管OnPro在Tiny-ImageNet上在记忆缓存器尺寸设为2 000时取得了最优的性能,但本文方法仅落后其0.48%. OnPro在带有更少任务数和更少类别数的小尺寸CIFAR-10数据集上展现出强劲的表现,但本文方法在带有更多任务和更多类别的大尺寸数据集CIFAR-100和Tiny-ImageNet上表现更优. 另外,表1中报告的结果展示出增加记忆缓存器尺寸可以增强所有比较方法的表现. 而本文方法可以从增大的记忆缓存器尺寸中受益更多. 例如,当记忆缓存器尺寸从1 000增加到5 000时,本文方法在CIFAR-100上的表现增加了18.67%,而第二大的增益由ER取得,为14.94%. 从结果中可以观察到,本文方法一致地取得了最大的准确率提升:在CIFAR-10上记忆缓存器尺寸从500增加到1 000时准确率增长了10.42%;在Tiny-ImageNet上,记忆缓存器尺寸从2 000增加到10 000时,准确率增长了17.2%.

表1　三种评估基准上,采用不同尺寸记忆缓存器时,本文方法和其他方法的平均最终准确率对比　　单位:%

数据集	CIFAR-10		CIFAR-100			Tiny-ImageNet		
比较方法	M=500	M=1 000	M=1 000	M=2 000	M=5 000	M=2 000	M=4 000	M=10 000
AGEM	22.7±1.9	22.6±0.7	5.8±0.2	5.8±0.3	6.5±0.2	0.9±0.1	2.1±0.1	3.9±0.2
GSS	30.7±1.3	40.1±1.4	11.1±0.2	13.3±0.5	17.4±0.1	3.3±0.5	10.0±0.2	10.5±0.2
MIR	40.0±0.6	41.0±0.6	15.7±0.2	19.1±0.1	24.1±0.2	6.1±0.5	11.7±0.2	13.5±0.2
ASER	36.2±1.2	44.7±1.2	16.4±0.3	12.2±1.9	27.1±0.3	5.3±0.3	8.2±0.2	10.3±0.4
GDumb	50.7±0.7	63.5±0.5	14.1±0.3	20.1±0.2	36.0±0.5	12.6±0.1	12.7±0.3	15.7±0.2
SCR	58.2±0.5	64.1±1.2	26.5±0.2	31.6±0.5	36.5±0.5	10.6±1.1	17.2±0.1	20.4±1.1
IL2A	56.0±0.4	58.2±1.2	18.2±1.2	19.7±1.5	22.4±0.2	5.5±0.7	8.1±1.2	11.6±0.4
Co^{2}L	51.0±0.7	58.8±0.4	17.1±0.4	24.2±0.2	32.2±0.5	10.1±0.2	15.8±0.4	22.5±1.2
OCM	70.0±1.3	77.2±0.5	28.1±0.3	35.0±0.4	42.4±0.5	15.7±0.2	21.2±0.4	27.0±0.3
ER	56.68±1.89	62.32±4.13	24.47±0.72	31.89±1.45	39.41±1.81	10.82±0.79	17.31±1.45	24.71±2.52
DER++	58.04±2.30	64.02±1.92	25.09±1.41	32.33±2.66	38.31±2.28	8.73±1.58	12.17±1.97	19.40±3.71
ERACE	53.26±3.04	59.95±2.40	28.36±1.99	34.21±1.53	39.39±1.31	13.56±1.00	19.00±0.87	25.92±1.07
GSA	60.34±1.97	66.54±2.28	27.72±1.57	35.08±1.37	41.41±1.65	12.44±1.17	18.45±1.07	25.34±1.43
OnPro	**70.47±2.12**	**74.70±1.51**	27.22±0.77	33.33±0.93	41.59±1.38	**14.32±1.40**	20.12±1.91	26.38±2.18
SSD	58.78±1.26	65.07±0.74	28.35±0.46	33.43±0.55	38.66±0.36	8.08±0.20	10.48±0.23	14.56±0.49
本文方法	58.90±2.65	69.32±0.83	**28.78±1.16**	**37.81±1.01**	**47.45±0.84**	13.84±0.92	**21.46±0.78**	**31.04±0.81**

注:加粗数据表示最优结果,下划线数据表示次优结果.

表2报告了不同方法的学习准确率. 这一度量反映了模型学习新任务的能力. 本文方法在相当大的程度上一致地超过了比较的方法:在CIFAR-10、CIFAR-100和Tiny-ImageNet上,在不同的记忆缓存器尺寸下,超过了比较的方法1.66%至70.4%. 从表2中可以观察到OnPro在小尺寸的CIFAR-10数据集上取得了第二优的学习准确率,仅仅落后本文方法4.73%(M=500)和2.96%(M=1 000). 而DER++在具有更长任务序列和更多类别的CIFAR-100和Tiny-ImageNet上展现了更强的学习新任务的能力. 虽然如此,本文方法体现出了实质性的优势,学习准确率超过它们1.66%至17.85%. 表2中结果体现本文方法在有效学习新任务方面的卓越能力.

表2 三种评估基准上,不同尺寸记忆缓存器下,本文方法和其他方法的学习准确率对比 单位:%

数据集	CIFAR-10		CIFAR-100			Tiny-ImageNet		
比较方法	M=500	M=1 000	M=1 000	M=2 000	M=5 000	M=2 000	M=4 000	M=10 000
ER	83.13±1.60	78.15±3.60	53.77±1.51	51.53±1.66	50.79±0.71	68.15±1.47	64.83±1.35	64.44±1.45
DER++	77.14±2.96	78.02±2.16	56.13±3.75	55.33±3.26	56.32±3.44	70.01±1.83	55.51±2.53	70.28±2.42
ERACE	57.66±4.16	61.59±3.35	38.53±1.61	39.95±2.00	41.56±1.44	5.60±1.45	4.83±0.74	4.92±0.95
GSA	79.87±3.26	77.09±4.55	58.16±1.58	55.13±1.81	50.34±1.73	20.46±1.59	17.52±1.00	14.50±0.63
OnPro	84.23±2.00	85.60±1.56	41.34±1.63	42.59±1.65	42.92±1.00	20.84±1.47	16.84±1.13	15.82±1.04
本文方法	**88.96±1.87**	**88.55±1.36**	**63.45±1.17**	**62.47±0.84**	**61.43±0.93**	**76.00±1.10**	**73.36±1.14**	**71.94±1.04**

注:加粗数据表示最优结果,下划线数据表示次优结果.

4.3 消去实验与分析

4.3.1 组件消去实验

为证实特征融合和一致性损失的有效性,本文在CIFAR-10和CIFAR-100上主导了消去实验,相关结果报告于表3,可以观察到,在基线方法上增加特征融合技术能提升平均最终准确率. 大多数情况下,在此基础上应用一致性损失可以进一步提升表现. 另外,特征融合和一致性损失的表现增益会随着记忆缓存器的增大而增加. 这些结果证实了本文方法每个组件的有效性. 值得注意的是,在小数据集CIFAR-10上,尽管一致性损失在存储样本数较少(M=500)时带来了轻微的负增益,但本文方法整体上依旧有效.

表3 本文方法的不同组件的消去实验 单位:%

方法组件			CIFAR-10		CIFAR-100		
Base	FF	CL	M=500	M=1 000	M=1 000	M=2 000	M=5 000
√			56.70±1.97	62.24±3.83	24.16±0.95	31.78±1.23	39.43±1.68
√	√		**59.52±1.58**	67.01±1.27	26.85±1.29	35.49±1.28	43.38±1.33
√	√	√	58.90±2.65	**69.32±0.83**	**28.78±1.16**	**37.81±1.01**	**47.45±0.84**

注:"Base"代表基线方法ER;"FF"代表特征融合;"CL"代表一致性损失;加粗数据表示最优结果.

4.3.2 锚点特征数量的影响

为研究锚点特征数量的影响,本文在CIFAR-100数据集上在记忆缓存器尺寸设为1 000的情况下做了实验. 表4展示了锚点特征数量分别设置为2、3、4、5时模型的平均最终准确率. 观察到,模型的表现对锚点特征的数量不是很敏感,因为准确率的变动仅仅0.35%. 当锚点特征数量选取为3时,模型可以取得相对更好的表现. 因此,在本文的其他实验中将锚点特征数量设置为3.

表4 锚点特征数量的影响探究 单位:%

m	2	3	4	5
本文方法	28.53±0.82	**28.78±1.16**	28.68±0.84	28.43±1.76

注:m为锚点特征数量;加粗数据表示最优结果.

4.3.3 平衡参数λ的影响

为分析一致性损失对模型性能的影响,将式(12)中的平衡因子λ分别设置为$[0.0, 0.1, 0.5, 1.0, 5.0, 10.0]$. 该实验在CIFAR-100上,在1 000的记忆缓存器尺寸下进行,相关结果报告于表5. 从表5中可知,运用一致性损失即使在小平衡因子λ=0.1下也能提升1.02%的表现. 这些预定义的λ取值中,λ=1.0能提升最多的平均最终准确率. 因此,本文在其他实验中将平衡因子默认设置为1.0.

表5 总损失中平衡因子λ对平均最终准确率的影响 单位:%

λ	0.0	0.1	0.5	1	5	10
本文方法	26.85±1.29	27.87±1.25	28.22±0.94	**28.78±1.16**	28.59±0.96	27.78±1.07

注:加粗数据表示最优结果.

4.3.4 小存储容量的影响

本文在CIFAR-100数据集上做了实验,研究本文方法在小尺寸记忆缓存器设置下的表现. 图3描绘了记忆缓存器大小设为100、200、300、400和500时的结果. 从图中可知,本文方法一致地超过了基线方法ER,证实了本文方法在使用小尺寸记忆缓存器时亦能有效工作. 另外,从图3中也可观察到对于本文方法和基线方法而言,增大记忆缓存器的尺寸都可以有效提升模型的平均最终准确率.

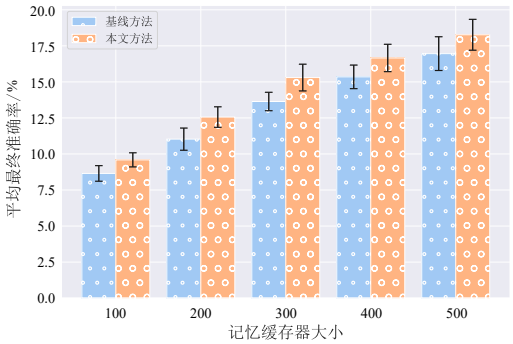


图3　在小尺寸记忆缓存器上的平均最终准确率的对比

4.3.5 枢纽特征是否增强的影响

为探究枢纽特征采用原始图片特征或是增强图片特征的影响,本文在CIFAR-10和CIFAR-100数据集上进行了实验,实验结果如下表6所示. 从表6中可知,在同基线方法保持一致的情况下,无论枢纽特征从原始图片提取还是从增强图片提取,本文方法均能够有效提高训练所得模型的平均最终准确率. 另外,亦可观察到,在使用原始图片提取枢纽特征时,本文方法带来的性能增益相比使用增强图片时更高.

表6　枢纽特征增强前后的平均最终准确率　　单位:%

图片	方法	CIFAR-10 (M=500)	CIFAR-100 (M=1 000)
原始	基线方法	31.30±3.95	11.54±0.40
	本文方法	43.41±2.84	26.85±1.29
增强	基线方法	56.68±1.89	24.47±0.72
	本文方法	58.90±2.65	28.78±1.16

4.3.6 不同的基线方法

本文方法和部分现有的在线连续学习方法互为补充. 为证实这一点,我们将所提方法运用于ER[8]、ERACE[12]和DER++[13]上. 表7展示了在CIFAR-10和CIFAR-100数据集上,在不同的记忆缓存器大小下和不同的基线方法基础上使用本文方法后,训练所得模型的平均最终准确率. 由表7可知,本文方法能有效提升基线方法的表现:在CIFAR-10上提升了1.43%~7.43%,在CIFAR-100上提升了3.04%~8.02%. 另外,由表7可知,本文方法在更大的记忆缓存器大小下,对基线方法的提升更为明显,例如CIFAR-10上M=500时,本文方法相比基线方法ER提升2.2%,但在M=1 000时,使用本文方法的提升扩大到7.08%.

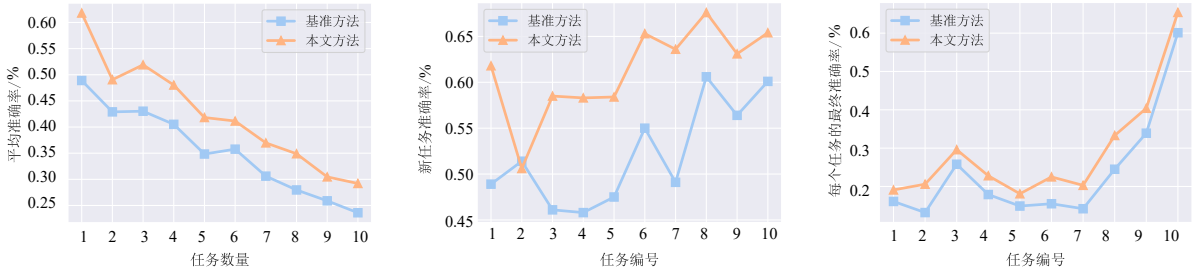
表7　本文方法运用于不同基线方法的表现　　单位:%

数据集	CIFAR-10		CIFAR-100		
方法	M=500	M=1 000	M=1 000	M=2 000	M=5 000
ER	56.70	62.24	24.16	31.78	39.43
+Ours	**58.90**	**69.32**	**28.78**	**37.81**	**47.45**
ERACE	53.58	60.40	28.30	34.49	39.82
+Ours	**60.27**	**67.83**	**31.36**	**38.19**	**46.66**
DER++	58.06	63.53	26.02	32.73	38.59
+Ours	**59.49**	**67.32**	**29.06**	**38.12**	**46.14**

注:加粗数据表示最优结果.

4.3.7 准确率的详细比较

为详细比较本文方法和基线方法随任务数增长的表现和在各个任务上的表现,我们在图4中绘制了多种准确率结果. 图4(a)描绘了随任务数量增加,所有已见任务的平均准确率. 从图中可知,本文方法在任意任务步上都实质性地超过了基线方法. 图4(b)展示了每个新任务首次被学习后,模型在新任务测试集上的表现. 可以看出,本文方法相比基线方法在几乎所有任务上都有更优的表现. 我们在图4(c)中展示了学习完最后一个任务后,每个任务的准确率. 可以看到,本文方法在所有任务上一致地超过了基线方法. 另外,图4(a)中平均准确率的下降和图4(c)中前9个任务的平均准确率明显低于当前任务的平均准确率的现象,反映了本文方法和基线方法在缓解灾难性遗忘方面能力的不足. 需指出的是,本文方法关注的目标是从数据流中有效地学习新知识,这是在线连续学习相比离线连续学习特有的关键挑战. 因此,如表2所示,在描述模型学习新任务能力的学习准确率指标下,本文方法的表现相比其他一些关注于缓解遗忘的方法要好很多.



(*a*) 已学习任务的平均准确率 (*b*) 新任务准确率 (*c*) 每个任务的最终准确率

图4 本文方法和基线方法间随任务数增长的准确率对比和各个任务上的准确率对比

4.3.8 基础训练超参数的影响

为详细探究基础训练超参数的选择对本文方法训练所得模型的影响,我们改变了新数据批大小、记忆数据重放批大小和学习率,在CIFAR-10(M=500)和CIFAR-100(M=1 000)数据集上进行了实验. 表8、表9和表10分别展示了新数据批大小、记忆数据重放批大小和学习率对平均最终准确率的影响. 从表8可知,本文方法受新数据批大小的影响不明显. 新数据批大小从5增长到30的过程中,在CIFAR-10和CIFAR-100数据集上的平均最终准确率的变化幅度均只在3%左右. 这是因为新数据会被选择性地存储到记忆缓存器中,即使当前新数据批大小很小,这部分数据也可能在后续训练过程中被重复使用. 从表9中可知,记忆数据重放批大小对训练所得模型的分类表现有较大影响. 随着重放批大小从8增大到128,平均最终准确率在CIFAR-10上的变化幅度有6%,在CIFAR-100上的变化幅度则达10%. 当重放批大小设置为最小值8时,训练所得模型取得了明显更低的性能. 从表10可知,学习率对训练所得模型的表现也有较大影响. 当学习率较大时(设为0.5),模型不能训练至收敛,其平均最终准确率仅为随机猜测的水平,在10类的CIFAR-10数据集上为10%,在100类的CIFAR-100数据集上为1%. 而当学习率较小时(设为0.000 05),模型的表现未能出现严重的衰退.

表8 新数据批大小对本文方法平均最终准确率的影响 单位:%

新数据批大小	5	10	15	20	30
CIFAR-10	60.30±1.68	58.90±2.65	60.93±2.04	60.05±2.34	57.92±1.11
CIFAR-100	27.98±0.61	28.78±1.16	27.67±1.14	26.82±1.31	25.19±0.98

表9 记忆数据重放批大小对本文方法平均最终准确率的影响 单位:%

重放批大小	8	16	32	64	128
CIFAR-10	56.28±3.17	59.92±2.07	62.25±1.40	58.90±2.65	59.86±3.08
CIFAR-100	18.38±1.17	25.01±0.95	27.68±1.06	28.78±1.16	27.80±1.05

表10 学习率对本文方法平均最终准确率的影响 单位:%

学习率	0.000 05	0.000 5	0.005	0.05	0.5
CIFAR-10	55.51±1.34	58.90±2.65	56.41±1.77	55.35±2.64	10.00±0.00
CIFAR-100	25.26±0.68	28.78±1.16	24.61±1.02	21.25±1.08	1.00±0.00

4.3.9 各方法训练时间的比较

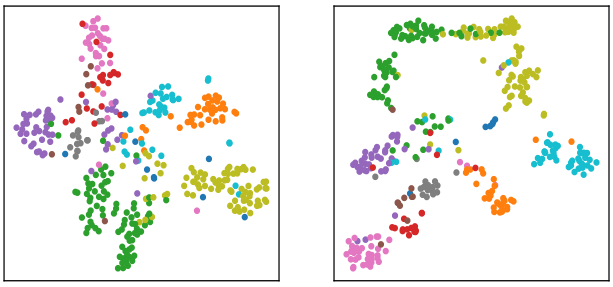
为比较不同方法的计算时间开销,我们报告了本文方法和主流方法的训练时间. 这些方法由于采取了同样的网络结构且是针对训练过程进行的设计,从而推断所需的时间是相近的,所以这里我们使用训练时间来衡量计算时间开销. 表11展示了不同方法在CIFAR-10(M=500)和CIFAR-100(M=1 000)上的训练时间. 从表中可知,本文方法取得了相比主流方法较为适中的计算开销. 其在CIFAR-10上的训练时间仅为OCM的42%,为OnPro的86%,在CIFAR-100上的训练时间仅OCM的39%,OnPro的52%. 而相比其余训练时间优于本文方法的算法,本文方法在CIFAR-10和CIFAR-100上均取得了更优的平均最终准确率,如表1所示.

表11 不同方法的训练时间比较 单位:s

方法	ER	DER++	ERACE	GSA	OCM	OnPro	本文方法
CIFAR-10	403±18	547±4	360±4	765±3	2 969±10	1 457±5	1 249±5
CIFAR-100	525±12	628±33	453±14	874±20	3 325±33	2 531±45	1 305±20

4.3.10 特征空间与融合权重可视化

为直观地比较本文方法和基线方法对特征空间的影响,我们绘制了本文方法和基线方法训练所得模型提取的CIFAR-10图片特征空间的t-SNE[37]可视化结果,如图5所示(图中每一种颜色代表同一种类别,每一个点代表一个样本特征). 从图中可以看出,本文方法使得模型提取的特征聚类更为集中,且不同类别之间的混淆更少. 另外,我们也绘制了在不同任务上训练完成后,本文方法对不同增强样本特征的融合权重谱,如图6所示(图中每一列代表三种数据增强方式,每一行代表不同的采样所得样本). 从图中可以观察到,对于不同的样本,每种数据增强产生的样本均有可能对融合样本有所贡献,本文方法不会使所有样本均偏向于某一种数据增强的方式.



(a) 基线方法　　(b) 本文方法

图5　模型提取的CIFAR-10图片特征可视化结果比较

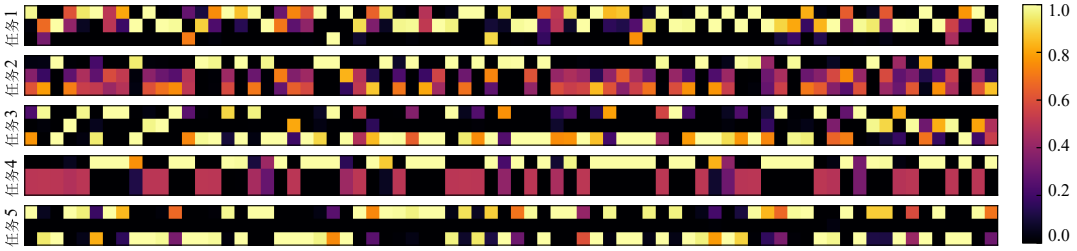


图6　本文方法对各种增强样本特征的融合权重谱

5　结束语

本文方法旨在处理在线类增量连续学习中有效学习新知识的问题. 为解决这一问题,我们提出一种新颖的特征融合方法,该方法利用了提取自增强样本的多个锚点特征,通过计算锚点特征和提取自输入图片或者增强图片的枢纽特征之间的相似度来加权融合这些锚点特征,以获取融合特征. 我们优化融合特征计算出的交叉熵损失,以促进新图片样本的学习过程. 另外,我们设计了一种一致性损失来加强约束融合特征和枢纽特征之间的距离,以进一步促进学习过程. 为充分地理解我们的方法,我们对模型参数的梯度进行了分析,阐述了特征融合与梯度重加权之间的关系. 实验结果证实了本文方法的有效性以及相比现有在线连续学习方法的竞争力. 本文方法关注于新知识的学习,在缓解灾难性遗忘方面存在不足. 后续工作中,我们将结合在线连续学习的特有性质,探索缓解遗忘并促进新知识学习的方法.

参考文献

[1] MA X Z, OUYANG W L, SIMONELLI A, et al. 3D object detection from images for autonomous driving: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3537-3556.

[2] 李博洋, 刘思健, 崔明月, 等. 基于最小回环检测的多车协同SLAM框架[J]. 电子学报, 2021, 49(11): 2241-2250.
LI B Y, LIU S J, CUI M Y, et al. Multi-vehicle collaborative SLAM framework for minimum loop detection[J]. Acta Electronica Sinica, 2021, 49(11): 2241-2250. (in Chinese)

[3] 周治国, 马文浩. 一种多层多模态融合3D目标检测方法[J]. 电子学报, 2024, 52(3): 696-708.
ZHOU Z G, MA W H. 3D object detection based on multilayer multimodal fusion[J]. Acta Electronica Sinica, 2024, 52(3): 696-708. (in Chinese)

[4] GAO C, LIU S, CHEN J Y, et al. Room-object entity prompting and reasoning for embodied referring expression [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 994-1010.

[5] 张磊, 鲁凯, 高春侠, 等. 基于变增益自抗扰技术的机器人轨迹跟踪控制方法[J]. 电子学报, 2022, 50(1): 89-97.
ZHANG L, LU K, GAO C X, et al. Path tracking control method of robot based on time-varying gain active disturbance rejection control[J]. Acta Electronica Sinica, 2022, 50(1): 89-97. (in Chinese)

[6] 金紫凤, 潘思聪, 危辉. 可变环境下基于位姿变换矩阵的机器人无标定手眼协调方法[J]. 电子学报, 2022, 50(10): 2318-2328.
JIN Z F, PAN S C, WEI H. Uncalibrated hand eye coordi-

nation method for robot based on pose transformation matrix in variable environment[J]. Acta Electronica Sinica, 2022, 50(10): 2318-2328. (in Chinese)

[7] VAN DE VEN G M, TUYTELAARS T, TOLIAS A S. Three types of incremental learning[J]. Nature Machine Intelligence, 2022, 4(12): 1185-1197.

[8] CHAUDHRY A, ROHRBACH M, ELHOSEINY M, et al. On tiny episodic memories in continual learning[EB/OL]. (2019-06-04) [2025-10-20]. https://arXiv.org/abs/1902.10486.

[9] CHAUDHRY A, RANZATO M, ROHRBACH M, et al. Efficient lifelong learning with A-GEM[EB/OL]. (2019-01-09)[2025-10-10]. https://arXiv.org/abs/1812.00420.

[10] PRABHU A, TORR P H S, DOKANIA P K. GDumb: A simple approach that questions our progress in continual learning[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 524-540.

[11] SHIM D, MAI Z D, JEONG J, et al. Online class-incremental continual learning with adversarial shapley value[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(11): 9630-9638.

[12] CACCIA L, ALJUNDI R, ASADI N, et al. New insights on reducing abrupt representation change in online continual learning[EB/OL]. (2022-05-02) [2025-10-10]. https://arXiv.org/abs/2104.05025.

[13] BUZZEGA P, BOSCHINI M, PORRELLO A, et al. Dark experience for general continual learning: A strong, simple baseline[EB/OL]. (2020-10-22) [2025-10-20]. https://arXiv.org/abs/2004.07211.

[14] MAI Z D, LI R W, KIM H, et al. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2021: 3584-3594.

[15] WEI Y J, YE J X, HUANG Z Z, et al. Online prototype learning for online continual learning[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 18718-18728.

[16] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[17] HOU S H, PAN X Y, LOY C C, et al. Learning a unified classifier incrementally via rebalancing[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 831-839.

[18] WU Y, CHEN Y P, WANG L J, et al. Large scale incremental learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 374-382.

[19] MITTAL S, GALESSO S, BROX T. Essentials for class incremental learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2021: 3508-3517.

[20] AHN H, KWAK J, LIM S, et al. SS-IL: Separated softmax for incremental learning[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 824-833.

[21] GUO Y D, LIU B, ZHAO D Y. Online continual learning through mutual information maximization[C]//International Conference on Machine Learning. Cambridge: PMLR, 2022: 8109-8126.

[22] CHA H, LEE J, SHIN J. Co2L: Contrastive continual learning[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 9496-9505.

[23] ZHU F, CHENG Z, ZHANG X Y, et al. Class-incremental learning via dual augmentation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 14306-14318.

[24] GU J Y, WANG K, JIANG W, et al. Summarizing stream data for memory-constrained online continual learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(11): 12217-12225.

[25] YU R N, LIU S H, WANG X C. Dataset distillation: A comprehensive review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 150-170.

[26] DE LANGE M, ALJUNDI R, MASANA M, et al. A continual learning survey: Defying forgetting in classification tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3366-3385.

[27] ALJUNDI R, LIN M, GOUJAUD B, et al. Gradient based sample selection for online continual learning[C]//Neural Information Processing Systems. Curran Associates Inc.: Red Hook, 2019: 11817-11826.

[28] ALJUNDI R, BELILOVSKY E, TUYTELAARS T, et al. Online continual learning with maximal interfered retrieval[EB/OL]. (2019-10-29)[2025-10-10]. https://arXiv.org/abs/1908.04742.

[29] GUO Y D, LIU B, ZHAO D Y. Dealing with cross-task class discrimination in online continual learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 11878-11887.

[30] WU Y C, WANG H, ZHAO P L, et al. Mitigating cata-

strophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization[C]// Proceedings of the 41st International Conference on Machine Learning. Cambridge: PMLR, 2024: 8929.

[31] HE Y H, CHEN Y J, JIN Y H, et al. DYSON: Dynamic feature space self-organization for online task-free class incremental learning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 23741-23751.

[32] SEO M, KOH H, JEUNG W, et al. Learning equi-angular representations for online continual learning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 23933-23942.

[33] PAPYAN V, HAN X Y, DONOHO D L. Prevalence of neural collapse during the terminal phase of deep learning training[J]. PNAS2020, 117(40): 24652-24663.

[34] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. (2009-04-08) [2025-10-10]. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[35] LE Y, YANG X. Tiny ImageNet visual recognition challenge[EB/OL]. (2015)[2025-10-10]. https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

[36] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[37] MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

附录A

在该附录中，我们对选取原始图片特征作为枢纽特征的原因做了理论分析，总结为下述定理.

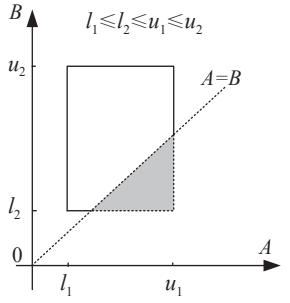
定理　在距离度量$D(\cdot,\cdot)$下，原始图片特征z与融合特征$\hat{z}$的距离$D(z,\hat{z})$不大于增强特征$\hat{z}_k$与融合特征$\hat{z}$之间距离$D(\hat{z}_k,\hat{z})$的概率多于0.5，即

$$P\left\{D(z,\hat{z})\leqslant D(\hat{z}_k,\hat{z})\right\}>0.5.$$

证明　将原始图片特征记为z，理想特征记为z^*，增强后图片特征记为$\hat{z}_k$，融合后特征记为$\hat{z}$. 选取原始图片特征z作为枢纽特征的出发点是该特征相比增强特征$\hat{z}_k$更靠近理想特征z^*，并且融合特征$\hat{z}$亦要比单一特征更靠近理想特征，可表示为$D(\hat{z},z^*)\leqslant D(z,z^*)\leqslant D(\hat{z}_k,z^*)$，其中$D(\cdot,\cdot)$表示两个特征间的距离.

首先分析$D(z,\hat{z})$和$D(\hat{z}_k,\hat{z})$的上下界. 由三角不等式可知，$D(z,\hat{z})\geqslant D(z,z^*)-D(\hat{z},z^*)\triangleq l_1$，且$D(\hat{z}_k,\hat{z})\geqslant D(\hat{z}_k,z^*)-D(\hat{z},z^*)\triangleq l_2$，又因为$D(z,z^*)\leqslant D(\hat{z}_k,z^*)$，所以$l_1=D(z,z^*)-D(\hat{z},z^*)\leqslant D(\hat{z}_k,z^*)-D(\hat{z},z^*)=l_2$. 类似地，由三角不等式亦可知，$D(z,\hat{z})\leqslant D(z,z^*)+D(\hat{z},z^*)\triangleq u_1$，且$D(\hat{z}_k,\hat{z})\leqslant D(\hat{z}_k,z^*)+D(\hat{z},z^*)\triangleq u_2$，又因为$D(z,z^*)\leqslant D(\hat{z}_k,z^*)$，所以$u_1=D(z,z^*)+D(\hat{z},z^*)\leqslant D(\hat{z}_k,z^*)+D(\hat{z},z^*)=u_2$. 综上可知，$l_1\leqslant l_2$，且$u_1\leqslant u_2$.

结合$D(\cdot,\cdot)\geqslant 0$的性质，容易得到$l_1\leqslant u_1$，且$l_2\leqslant u_2$，这样关于$l_1$、$l_2$、$u_1$和$u_2$的大小关系，有两种情况可以分析，下面分别讨论. 第一种情况，如果$l_1\leqslant u_1\leqslant l_2\leqslant u_2$，因为$l_1\leqslant D(z,\hat{z})\leqslant u_1$，且$l_2\leqslant D(\hat{z}_k,\hat{z})\leqslant u_2$，则必有$D(z,\hat{z})\leqslant D(\hat{z}_k,\hat{z})$；第二种情况，如果$l_1\leqslant l_2\leqslant u_1\leqslant u_2$，则$D(z,\hat{z})$和$D(\hat{z}_k,\hat{z})$所在的区间有重叠，可能出现$D(z,\hat{z})>D(\hat{z}_k,\hat{z})$的情形，下面分析出现这一情形的概率. 定义$A\triangleq D(z,\hat{z})$，$B\triangleq D(\hat{z}_k,\hat{z})$，该问题可以转化为如下问题：随机变量$A$均匀分布于区间$[l_1,u_1]$，随机变量$B$均匀分布于区间$[l_2,u_2]$，求$A>B$的概率. 从图A1中可知，$A>B$的概率为阴影部分面积同矩形面积之比，即$P(A>B)=\dfrac{0.5\times(u_1-l_2)\times(u_1-l_2)}{(u_1-l_1)\times(u_2-l_2)}$，结合$l_1\leqslant l_2\leqslant u_1\leqslant u_2$可知$P(A>B)\leqslant 0.5$. 综上可知，在多于0.5的概率下，有$D(z,\hat{z})\leqslant D(\hat{z}_k,\hat{z})$，因此选取原始图片特征作为枢纽特征.



图A1　图示法求解$A<B$的概率

作者简介

邱奔流 男,1999 年 5 月生,湖南常德人. 电子科技大学信息与通信工程专业博士研究生. 主要研究方向为连续学习、迁移学习、多模态学习、目标检测、社交网络分析.
E-mail: qbenliu@163.com

王岚晓 女,1997 年 3 月生,山东淄博人. 电子科技大学特聘副研究员. 主要研究方向为多媒体信息处理、多模态场景解析. 中国电子学会会员编号:E190091313M.
E-mail: lanxiaowang@foxmail.com

邱荷茜 女,1994 年 10 月生,山西运城人. 电子科技大学信息与通信工程学院副教授. 主要研究方向为计算机视觉、多媒体智能信息处理、目标检测与识别.
E-mail: hqqiu@uestc.edu.cn

高翔宇 男,1998 年 2 月生,贵州贵阳人. 电子科技大学信息与通信工程专业博士研究生. 主要研究方向为目标检测、开放词汇识别.
E-mail: xygao@std.uestc.edu.cn

问海涛 男,1995 年 4 月生,江苏淮安人. 电子科技大学信息与通信工程学院博士研究生. 主要研究方向为计算机视觉、机器学习、连续学习、增量学习.
E-mail: haitaowen@std.uestc.edu.cn

李宏亮 男,1970 年 8 月生,2005 年获西安交通大学博士学位,目前为电子科技大学二级教授、博士生导师,国家杰出青年科学基金获得者. 主要研究方向为多媒体智能、对象检测与分割、视觉感知模型以及机器学习等. 电子学会会员编号:E190008433S.
E-mail: hlli@uestc.edu.cn