

## 融合平衡权重和自监督的类增量学习方法

巩佳义, 许鑫磊, 肖婷, 王喆<sup>+</sup>

华东理工大学 信息科学与工程学院, 上海 200237

+ 通信作者 E-mail: wangzhe@ecust.edu.cn

**摘要:** 针对图像分类中类增量学习的知识灾难性遗忘现象, 现有类增量学习方法着重于模型分类层的不平衡偏移修正, 忽视了模型特征层的偏移, 未能良好解决类增量学习所面临的新旧类别不平衡问题。为此, 提出一种新的类增量学习方法, 称为融合平衡权重和自监督的类增量学习方法(BWSS)。首先, BWSS利用旧类在训练中预测期望低的特点设计了自适应的平衡权重, 扩大旧类的损失回传占比以修正整个模型的偏移。其次, BWSS引入自监督模块, 将样例的旋转角度预测作为辅助任务, 强化模型对冗余特征和共性特征的表达能力, 以更好地支撑增量任务。通过与主流类增量学习算法在公开数据集 CIFAR-10 和 CIFAR-100 上的实验对比, 证明 BWSS 不仅在类别少样例多的 CIFAR-10 上增量性能更优, 在类别多样本少的 CIFAR-100 上同样具有优势。通过消融实验和特征可视化, 验证了所提方法对模型的特征表示能力和增量性能是有效的。BWSS 在 CIFAR-10 上的 5 阶段增量任务最终准确率达到 76.9%, 相比基线方法提高了 5 个百分点。

**关键词:** 增量学习; 自监督; 不平衡问题; 知识蒸馏

**文献标志码:** A **中图分类号:** TP391

## Class Incremental Learning Method Integrating Balance Weight and Self-supervision

GONG Jiayi, XU Xinlei, XIAO Ting, WANG Zhe<sup>+</sup>

School of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

**Abstract:** In view of the catastrophic forgetting phenomenon of knowledge in class incremental learning in image classification, the existing class incremental learning methods focus on the correction of the unbalanced offset of the model classification layer, ignoring the offset of the model feature layer, and fail to solve the problem of the imbalance between the new and old samples faced by class incremental learning. Therefore, a new class incremental learning method is proposed, which is called balance weight and self-supervision (BWSS). BWSS designs an adaptive balance weight based on the low expectation of the old class in training, so as to expand the loss return proportion of the old class in the same data batch to correct the overall model offset. Then, BWSS introduces self-supervised learning to predict the rotation angle of the sample as an auxiliary task, so as to make the model have the expression ability of redundant features and common features to better support incremental tasks. Through the experimental comparison with the mainstream incremental class learning algorithms on the open datasets CIFAR-10 and CIFAR-100, it is proven that BWSS not only has better incremental performance on CIFAR-10 with fewer categories and more samples, but also has advantages on CIFAR-100 with more categories and fewer samples. Ablation experiments and fea-

**基金项目:** 上海市科技计划项目(21511100800); 国家自然科学基金(62076094); 国家部委科技领域基金。

This work was supported by the Science and Technology Program of Shanghai (21511100800), the National Natural Science Foundation of China (62076094), and the Program of Science and Technology of China Ministries and Commissions.

**收稿日期:** 2022-12-19 **修回日期:** 2023-03-21

ture visualization demonstrate that the proposed method is effective for the feature representation and incremental performance of the model. The final accuracy of BWSS's 5-stage incremental task on CIFAR-10 reaches 76.9%, which is 5 percentage points higher than the baseline method.

**Key words:** incremental learning; self-supervision; problem of imbalance; distillation of knowledge

在大数据时代背景下,海量训练数据往往以数据流的形式出现,这要求深度模型的训练需要在保留旧知识的前提下能较好地适应新的学习领域。为此,研究人员提出面向图像分类的类增量学习。类增量学习在数据隐私和数据存储领域具有良好前景和现实意义。在类增量学习的设置下,深度模型不再使用一个完整的数据集进行训练,而是把数据集划分到不同的任务。不同任务间的数据不互通或仅有少量数据可以在任务间流通。然而,在这样的设置下,模型学习新任务知识后,在旧任务上的表现会急剧下降,这一现象被称为灾难性遗忘<sup>[1]</sup>。

针对灾难性遗忘问题,文献[2]将旧模型作为教师模型,通过知识蒸馏<sup>[3]</sup>的方式让新模型的输出与旧模型对齐,以此保留旧知识。在此基础上,文献[4]使用自编码器获取每个任务中的关键信息,将特征投影至低维子流形中并传递至新模型。文献[5]通过对模型参数进行针对性约束来控制模型的知识保留和新知识学习。文献[6]和文献[7]通过训练生成式网络来为新任务回放旧样例,以此保持新旧类分类边界的平衡。上述都是基于无样例回放的类增量学习方法,模型不会保存任何旧任务中的样例数据。但即使是生成式网络也面临知识遗忘问题,因此无样例回放的方法在新旧类的判别性上表现不佳。基于样例回放的类增量学习方法允许使用有限的内存来保存旧任务中的部分数据。比如,文献[8]通过构建并管理一个样例集来保存部分旧数据,在训练阶段将其与新数据混合。由于内存有限,模型会在当前任务结束后剔除内存中部分旧数据并更新入新数据。文献[9]在此基础上通过约束旧样本在新旧模型下的特征余弦距离来维持旧任务的知识。在基于样例回放的情况下全连接分类层会对新类产生严重判别偏向,由此文献[10]和文献[11]分别从全连接分类层的输出结果和分类层的模型参数入手,扩大旧类的预测期望,以此平衡新旧类之间判别。

但上述方法忽视了特征提取器同样存在的偏移问题。图像分类深度模型主要由特征提取器和分类器组成,由于基于样例回放的方法保存的旧样例数

量相对于新样例过少,这使模型整体面临一个严重的不平衡问题。旧样例的特征空间在模型学习新任务时会不可避免地向新样例特征空间产生偏移,因而降低模型在新旧类之间的判别能力。为此,本文提出一种融合平衡权重和自监督(balanced weight and self-supervision, BWSS)的类增量学习方法。本文立足基于正则约束的类增量学习方法,并根据偏移修正的思想,从特征空间的修正和构建两个角度做出创新改进。首先,为了处理新旧类样例不平衡带来的模型偏向,本文利用旧类预测期望低会引起较大梯度这一特点,通过梯度生成权重,以此在训练时扩大旧样例在同一批样例中的损失回传,修正模型整体对新数据的偏向问题。其次,在模型构建特征空间时,本文引入旋转自监督辅助任务,利用自监督学习来提取样例中的冗余特征,构建鲁棒特征空间,从而达到缓解知识遗忘目的。

## 1 相关工作

### 1.1 基于正则约束的方法

很多研究人员就类增量学习中的知识灾难性遗忘问题做出了工作和贡献,基于正则约束的方法通过向损失函数中添加正则项来约束模型的学习过程,比如知识蒸馏项和参数约束项。EWC(elastic weight consolidation)<sup>[5]</sup>、MAS(memory aware synapses)<sup>[12]</sup>试图估计增量模型中各个参数对增量过程的重要性,并对重要参数的变化施加惩罚损失来进行约束。二者之间的差异在于计算参数重要性的方法。EWC通过估计Fisher Information Matrix的对角线值来计算参数重要性。与EWC不同的是,MAS使用在线学习的方法,基于样本的预测输出对参数变化敏感度进行估计,以此进一步测量网络的每一个参数重要性。

然而,设计一个合理的度量来评估所有参数是比较困难的,尤其是在长序列的任务或类别中。UCIR(unified classifier incrementally via rebalancing)<sup>[9]</sup>和PODNet<sup>[13]</sup>在模型的蒸馏损失上进行改进。UCIR使用余弦归一化强行平衡新旧类的幅值,并引入特征

蒸馏,而非传统的 logits 蒸馏,以此来约束模型在增量过程中对旧知识的遗忘。PODNet 在特征蒸馏的基础上,提出在特征提取器各级间进行蒸馏的想法,即对中间层的不同维度进行池化并形成不同强度的蒸馏。

### 1.2 基于偏移修正的方法

基于偏移修正的方法旨在解决增量过程中模型对新类的偏向问题。Bic<sup>[10]</sup>将混用的新旧类数据分成训练集和验证集两部分,训练集用于训练当前模型,而验证集用来对最后全连接层纠正。纠正的方式为保持训练模型不变,对模型最后的输出结果进行线性拟合修正。WA<sup>[11]</sup>从全连接分类层的权重出发,通过将新类权重向量与旧类权重向量进行范数对齐来进行全连接层的修正。PASS<sup>[14]</sup>作为无样例回放的方法,为了避免新类占据旧类分类边界而产生的分类偏移,提出原型增强的方法,保存旧类的原型并通过在分类层回放高斯噪声加强的原型来保持旧类的分类边界。

## 2 融合平衡权重和自监督的类增量学习

### 2.1 问题描述

类增量学习的目标是持续地学习一个统一的模型,以对目前接触过的所有类别的测试样本进行分类,而这些类别是分任务到达的。假设增量任务序列为  $\{t_0, t_1, \dots, t_n\}$ , 其中  $t_i$  的训练集为  $D_i$ , 测试集为  $E_i$ 。当模型  $\theta$  进行到任务  $t_i$  时,可供  $\theta$  训练用的当前数据集  $D_{cur} = D_i \cup D_{old}$ ,  $D_{old}$  为样例内存中的旧数据,包含了旧任务  $\{D_0, D_1, \dots, D_{i-1}\}$  中的极少部分样例。模型  $\theta$  在任务  $t_i$  的学习目标为最大化当前总测试集  $\{E_0, E_1, \dots, E_i\}$  的准确率。值得注意的是,当前任务模

型可以在旧任务模型的基础上进行训练,或者使用旧模型进行知识传递。但在测试时,不可以引用旧模型或旧模型的分层器来针对旧任务测试集进行测试,即类增量学习的结果只有一个最终模型。

### 2.2 总体框架

为构建鲁棒的特征空间并在增量过程中进行有效的特征修正,本文提出融合平衡权重和自监督的类增量学习方法,如图 1 所示。本模型在基于样例回放的类增量学习基础上引入自适应的平衡权重模块,在模型训练时,通过扩大作为少数样本的旧样例梯度回传来修正模型整体趋向新数据的偏移问题。而且模型在整体框架上额外引入样例旋转角度预测任务来帮助模型提取冗余信息。模型通过伪标签蒸馏法来进行知识传递。

### 2.3 平衡权重

在样例回放的类增量学习背景下,有限内存保存的少量旧数据样本和大量新数据样本一起用于新模型的训练。这时新模型面临的是一个新旧样例数据不平衡的局面,相对占多数的新样例会使模型参数偏向于新样例。为合理调整二者对模型参数的影响,本文提出平衡权重(balance weight, BW)方法。

给定当前任务  $t$  的模型参数  $\theta^{[t]}$ , 数据集  $D_{cur}$ , 则分类损失函数为:

$$L_{ce} = \frac{1}{S} \sum_{i=1}^S B_{ce}(P^{[t]}(x_i^{[t]}, \theta^{[t]}), y_i^{[t]}) \quad (1)$$

其中,  $S$  为数据批次大小(batchsize),  $x_i^{[t]}$  为样本数据,  $y_i^{[t]}$  为样本标签,  $(x_i^{[t]}, y_i^{[t]}) \in D_{cur}$ , 实际使用时单个样例  $(x_i^{[t]}, y_i^{[t]})$  的标签表现形式为 one-hot 向量  $y_i^{[t]}$ 。  $P^{[t]}(x_i^{[t]}, \theta^{[t]})$  为模型  $\theta^{[t]}$  的 softmax 输出期望。  $B_{ce}$  为二元交叉熵损失。

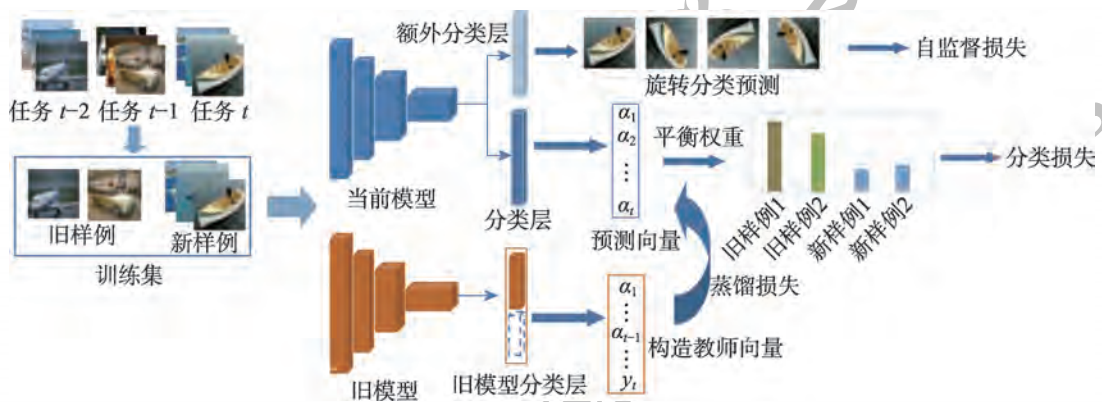


图1 融合平衡权重和自监督学习的类增量学习框架

Fig.1 Framework of class-incremental learning with balanced weights and self-supervision

通过对损失函数求导可得样例  $i$  在回传经过最后全连接层的梯度为:

$$g_i = \frac{\partial B_{CE}(P^{[i]}(x_i^{[i]}, \theta^{[i]}), y_i^{[i]})}{\partial \theta^{[i]}} = P^{[i]}(x_i^{[i]}, \theta^{[i]}) - 1 \quad (2)$$

根据文献[15]可知,作为不平衡数据中的少数类,旧样例往往会有较小的期望值,又因为  $P^{[i]}(x_i^{[i]}, \theta^{[i]}) < 1$ ,由此可知旧样例对应的梯度绝对值  $|g_{old}^{[i]}|$  要大于新样例对应的梯度绝对值  $|g_{new}^{[i]}|$ 。基于此,引入自适应的平衡权重的分类损失为:

$$L_{WC} = \frac{1}{S} \sum_{i=1}^S w_i \cdot B_{CE}(P^{[i]}(x_i^{[i]}, \theta^{[i]}), y_i^{[i]}) \quad (3)$$

其中,  $w_i = \frac{|g_i^{[i]}|}{|g^{[i]}|}$ ,  $|g^{[i]}|$  为当前数据批次中各样例对应的梯度平均值。

平衡权重的引入可以最大化发挥内存中旧样例的修正作用。且学习新样例的过程中直接在损失函数上提高旧样例的比重,可以在梯度回传时影响整个模型,而非仅仅最后的分类层,以此有效修正由于新旧数据数量不平衡而引起的模型偏移问题。

## 2.4 自监督辅助任务

在数据量较小或者类别较单一的情况下,简单分类任务易使模型关注局部易于分类的明显特征,比如分类辣椒和香蕉时使用颜色就可以良好分类。而类增量学习把完整的数据集分入各子任务中,便使得模型面临上述情况。为充分利用各增量任务内的有限数据,本文引入旋转自监督辅助任务(self-supervised learning pretext task, SSL)来提取冗余特征。

受 RotationSSL<sup>[16]</sup>启发,首先分别对每个样例进行  $90^\circ$ 、 $180^\circ$ 、 $270^\circ$  的旋转并将旋转角度作为样例的增广标签,得到的构造样例为:

$$x_{r_i}^{[i]} = rotate(x_i^{[i]}, a), a \in \{0, 90, 180, 270\} \quad (4)$$

然后通过在原模型的基础上额外添加一个旋转分类层来完成旋转角度预测的辅助任务,旋转自监督损失为:

$$L_{SSL} = \frac{1}{4S} \sum_{i=1}^{4S} B_{CE}(P^{[i]}(x_{r_i}^{[i]}, \theta^{[i]}), y_{r_i}^{[i]}) \quad (5)$$

自监督辅助任务可以促使模型关注冗余特征来弥补简单分类任务的缺陷,比如形状、大小、相对位置等。冗余特征往往是各个增量任务中数据所共有的一些特征,因此使得模型具有更好的可塑性。模型训练过程中,由于训练数据进行了4倍扩充,产生了较多的时间花费,但增量学习的主要目标在于缓解

知识遗忘问题,可控的时间花费是可接受且值得的。

## 2.5 总损失函数

本节介绍本文模型训练时所使用的总损失函数。为了有效传递增量过程中的知识,本文使用了标签蒸馏损失。蒸馏方法为使用旧模型作为教师模型,通过模型输出对齐来实现知识传递。假设增量进行到任务  $t$  时,样例类别数有  $n_t$ ,当前模型最后全连接层对样例  $i$  的预测为  $p_i^{[t]} = (v_0, v_1, \dots, v_{n_t})_i$ ,旧模型最后全连接层预测输出为  $p_i^{[t-1]} = (v_0, v_1, \dots, v_{n_{t-1}})_i$ 。由于当前模型的分层需要在旧模型的基础上延伸出新类的预测神经元,这使得  $n_t$  要大于  $n_{t-1}$ 。为弥补二者长度差异,本文将样例  $i$  的 one-hot 标签填入  $p_i^{[t-1]}$  中缺失部分,得到新的教师模型预测向量,即:

$$t_i^{[t-1]} = p_i^{[t-1]} \cup (y_{n_{t-1}+1}, y_{n_{t-1}+2}, \dots, y_{n_t})_i \quad (6)$$

将此作为教师输出对新模型进行传递约束,可得蒸馏损失为:

$$L_{KD} = D_{KL}(p_i^{[t]} || t_i^{[t-1]}) \quad (7)$$

其中,  $D_{KL}$  为 KL 散度。

将上述各项结合后,模型的总损失函数为:

$$L_{total} = L_{WC} + \eta_1 \times L_{SSL} + \eta_2 \times L_{KD} \quad (8)$$

其中,  $\eta_1$ 、 $\eta_2$  为旋转自监督和蒸馏损失的系数。本文将  $\eta_1$ 、 $\eta_2$  分别设置为 1.5 和 3。

## 3 实验及分析

### 3.1 实验数据集以及场景划分

为验证本文模型 BWSS 的性能,实验所采用的数据集为 CIFAR-100 和 CIFAR-10 数据集。CIFAR-100 和 CIFAR-10 是一个包含 8 000 万张图像的数据集的带标签子集,如图 2 所示。

CIFAR-100 数据集总共有 100 个类别,每个类包含 600 张像素为  $32 \times 32$  的彩色图像,其中包含 500 张训练图像和 100 张测试用图像。CIFAR-10 数据集总共有 10 个类别,每个类别包含 6 000 张像素为  $32 \times 32$  的彩色图像,其中包含 5 000 张训练图像和 1 000 张测试图像。相比于 CIFAR-100, CIFAR-10 具有总类别少、单类数据量大的特点。为进行增量学习实验,本文将 CIFAR-100 和 CIFAR-10 分别划分为 5 阶段增量和 10 阶段增量。对于 CIFAR-100 数据集,随机选取 10 个类作为基数据集,随后将剩余的 90 个类 10 个为 1 组作为增量任务。对于 CIFAR-10 数据集,随机选取 2 个类作为基数据集,随后将剩余的 8 个类 2 个为 1 组作为增量任务。



图2 CIFAR 数据集示例

Fig.2 Example of CIFAR dataset

### 3.2 实验设置及评价指标

本文实验在 Red Hat 8.2.1-3 系统上进行, GPU 为单块 RTX-3090, 实验环境为 CUDA 11.4+python 3.6 + torch1.7.0。骨干网络为 Resnet-18<sup>[17]</sup>。在模型的训练参数设置中, batchsize 设置为 128, 样例回放内存大小设置为 2 000, 各旧类均分内存空间。模型训练使用 Adam 优化器, 初始学习率为 0.001, 每轮增量任务训练 100 轮, 每 45 轮学习率衰减为原 10%。为了保证实验的严谨性和公平对比, 本文所使用的运行环境、骨干网络、优化器设置和样例内存均与各对比方法保持一致。

本文评价指标为准确率  $Acc$ , 为模型对测试集图像类别的正确预测数占测试集图像总数的百分比, 即:

$$Acc = \frac{|E_{correct}|}{|E_{total}|} \times 100\% \quad (9)$$

其中,  $E_{total}$  指当前增量任务测试集以及之前所有增量任务测试集之和。

### 3.3 对比模型简介

本文实验对比了主流类增量学习算法, 包括 LwF、iCaRL、Bic、WA 和 PODNet。iCaRL、Bic、PODNet 和 WA 作为有样例回放的类增量学习方法, 都使用了内存来保存部分旧样例。其中, Bic 通过线性拟合修正全连接层分类器的分类偏好来提高准确率, 而 WA 修正全连接层偏好的方式为权重对齐。LwF 和 EWC 为无样例回放的类增量学习方法。EWC 在 LwF 的基础上分别进行了参数约束和知识蒸馏的改进。iCaRL 为本模型的基模型。

### 3.4 对比实验

本文所提模型 BWSS 与上述对比模型的实验结果如表 1 和表 2 所示。表中展示了各模型在 CIFAR-

表1 CIFAR-100 数据集增量对比结果

Table 1 Incremental comparison results on

增量模型	基数数据集	CIFAR-100 dataset								
		任务								
LwF	90.2	69.8	57.3	43.6	39.3	32.1	30.2	25.2	24.6	23.3
EWC	90.2	32.1	33.9	24.3	24.2	19.4	18.8	16.3	15.1	12.5
PODNet	89.6	72.0	65.7	57.5	53.6	49.6	45.9	42.1	39.3	36.5
iCaRL	89.6	75.6	69.4	65.6	61.9	59.5	57.5	55.0	53.3	50.3
Bic	90.3	76.5	72.7	67.8	64.6	61.8	59.0	55.6	52.5	50.6
WA	90.4	77.1	73.7	68.2	65.6	61.7	58.8	54.9	52.8	50.5
<b>BWSS</b>	<b>91.3</b>	<b>82.4</b>	<b>76.5</b>	<b>70.4</b>	<b>66.4</b>	<b>62.8</b>	<b>60.4</b>	<b>55.9</b>	<b>54.6</b>	<b>51.2</b>

表2 CIFAR-10 数据集上增量对比结果

Table 2 Incremental comparison results

增量模型	基数数据集	on CIFAR-10 dataset			
		任务			
LwF	92.5	70.8	55.2	40.9	40.3
EWC	93.1	49.1	32.0	24.7	10.9
PODNet	94.5	86.6	78.0	67.9	68.6
iCaRL	92.9	87.7	78.2	69.5	71.9
Bic	92.6	87.3	81.1	73.0	70.5
WA	92.9	86.5	80.5	70.7	74.5
<b>BWSS</b>	<b>94.9</b>	<b>90.0</b>	<b>81.9</b>	<b>73.4</b>	<b>76.9</b>

100 和 CIFAR-10 数据集上各增量阶段的分类准确率。

从表 1 可知, BWSS 模型的增量实验结果优于其他对比模型。BWSS 在 CIFAR-100 中的基数数据集准确率达到 91.3%, 增量任务 1 准确率达到 82.4%, 相较于 LwF 分别提升了 1.1 个百分点与 12.6 个百分点。这证明即使只使用有限内存保存少数样例带来的优势也是巨大的, 而且随着增量任务的进行, 这一优势呈逐步扩大趋势, 在最终增量任务 9 上 BWSS 比 LwF 提高了 27.9 个百分点。BWSS 相较于基模型 iCaRL 的基数数据集和增量任务 1 分别提升了 1.7 个百分点与 6.8 个百分点, 这一差距在任务 2 上达到最大值 7.1 个百分点, 且在随后增量任务中都保持了优势。这证明 BWSS 不单是改进了基模型准确率, 更重要的是模型具有更优秀的抗遗忘能力, 使得模型在增量开始阶段能够扩大差距。

表 2 展示了 BWSS 与各模型在 CIFAR-10 数据集上的对比实验。BWSS 在基数数据集上准确率达到 94.9%, 任务 1 的准确率达到 90.0%, 达到对比模型中最优的结果, 且比基模型 iCaRL 的基数数据集和任务 1 准确率分别高出 2.0 个百分点和 2.3 个百分点, 这一

差距呈逐步扩大趋势,在任务4上的差距达到了5.0个百分点。而相比于WA, BWSS在各任务上的优势分别为2.0个百分点、3.5个百分点、1.4个百分点、2.7个百分点和2.4个百分点,保持了全面的优势。从数据集间的对比来看,由于CIFAR-10数据集具有类别少、单个类别样例数多的特点,回放内存可以保存相对较多的旧样例,这使得CIFAR-10数据集的实验结果普遍高于CIFAR-100数据集的同期结果。

### 3.5 消融实验

本文对所提创新点平衡权重和旋转自监督辅助任务在CIFAR-100数据集上进行消融实验,实验结果如图3所示。蓝色虚线展示了基模型iCaRL实验结果,橙色实线展示仅使用平衡权重(BW)的实验结果,绿色实线为仅添加旋转自监督辅助任务(SSL)的实验结果,红色为BWSS的实验结果。为方便对比同为偏移修正方法的Bic和WA,本文在图中添加了代表Bic的红色虚线和代表WA的紫色虚线。

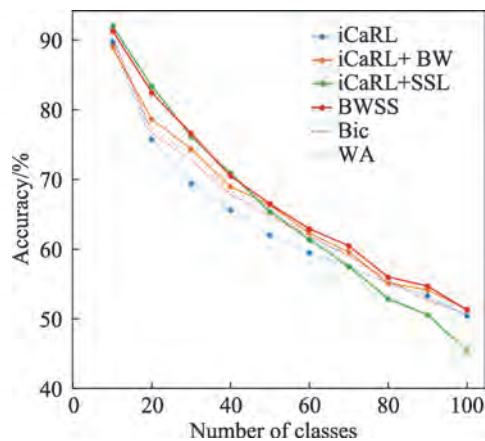


图3 消融实验对比

Fig.3 Comparison of ablation experiments

由图3可以看出,仅使用样例平衡权重的基模型实验结果好于WA和Bic。WA和Bic的修正方法仅影响全连接分类层而无法修正特征提取器上的特征偏移,而样例平衡权重从损失回传角度出发,有效发挥旧样例修正整体模型偏移的作用。在仅使用自监督辅助任务的情况下,模型在增量初期保持了较好的准确率,但在后期性能下降严重。造成这一现象主要原因在于当增量任务逐渐增加时,样例间特征差异逐渐扩大,且由于自监督辅助任务为模型带来较强的特征表示能力,特征偏移现象会更加严重。而初期较好的实验结果是由于在样例较少的情况下,样例间共有的部分鲁棒特征缓和了特征偏移现

象。因此通过融合平衡权重来修正特征偏移可以达到最优的结果。

### 3.6 平衡权重分析

本节在CIFAR-10数据集的各增量任务中提取新旧类的平衡权重并取平均进行分析。如图4所示,相比于新类,旧类始终保持了一个较大的权重,且呈逐渐扩大趋势。这些自适应的平衡权重将提高同一批次中旧任务样例的损失回传占比。与WA和Bic相比,通过扩大损失回传占比的方式不仅可以让旧样例在分类层发挥作用,还可以加强旧样例对特征提取器的修正作用,以此充分发挥旧类在增量任务中抗遗忘的作用。

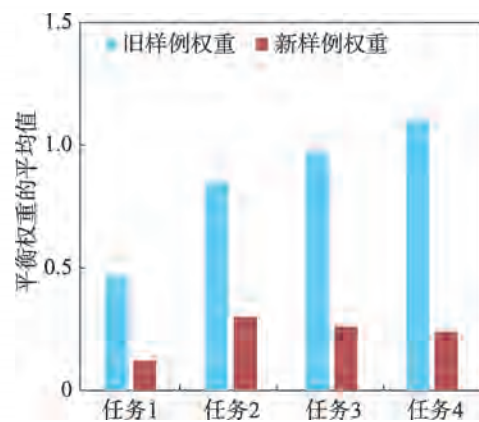


图4 平衡权重均值图

Fig.4 Average value of balance weight

### 3.7 特征空间可视化

为验证本文方法BWSS对模型特征提取能力的提升,本文通过T-SNE(t-distributed stochastic neighbor embedding)对CIFAR-10数据集的样例特征空间进行可视化展示并与基模型结果对比。图5(a)为基数据集在基模型的初始特征分布,图5(b)为基数据集在增量任务1结束后的基模型特征分布。同理,图5(c)为基数据集在BWSS上的初始分布,图5(d)为基数据集在BWSS的增量任务1结束后的特征分布。

由图5可见,在基数据集上,BWSS的特征空间分布相比于基模型分布更加均匀,而相比之下基于普通分类方法的基模型出现过拟合现象,类内样例分布过于集中。当进行到增量任务1时,基模型的类间样例特征分布出现靠拢现象,且部分样例完全混合。BWSS由于自监督学习的引入,使得模型具有冗余特征和共性特征表达能力。这些特征不仅能够运用在旧任务数据集,还对新任务数据集具有一定表达能力。且平衡权重通过适应性调整旧样例在分

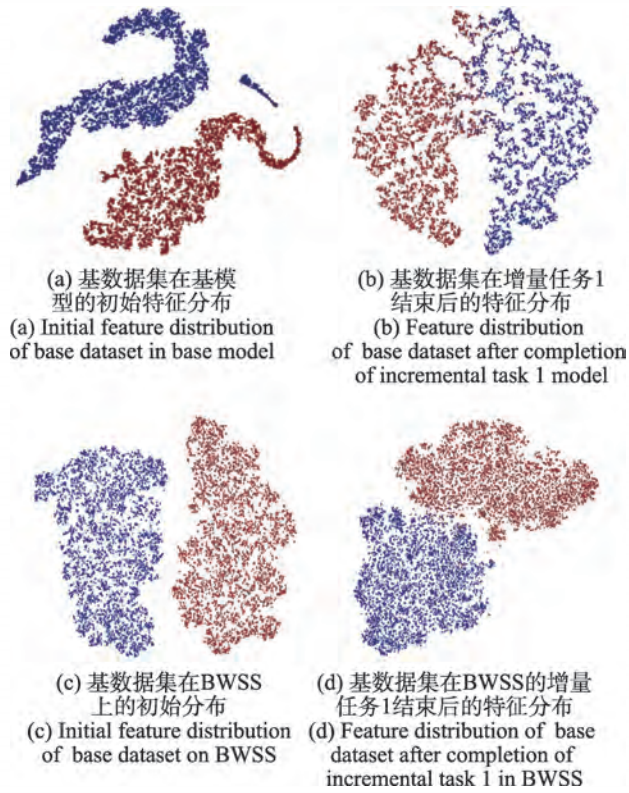


图5 特征空间可视化图

Fig.5 Visualization of feature space

类损失中的回传占比,有效避免旧样例的特征空间偏移,保持了更好的分类边界。因此,样例在BWSS上能够得到更加清晰的特征空间分布。

### 3.8 收敛性分析

本节展示了BWSS在CIFAR-100数据集的训练过程中各增量任务的准确率的收敛曲线,如图6所示。BWSS总损失函数包含分类损失、自监督辅助任务损失和蒸馏损失。在整体训练过程中,为了清晰

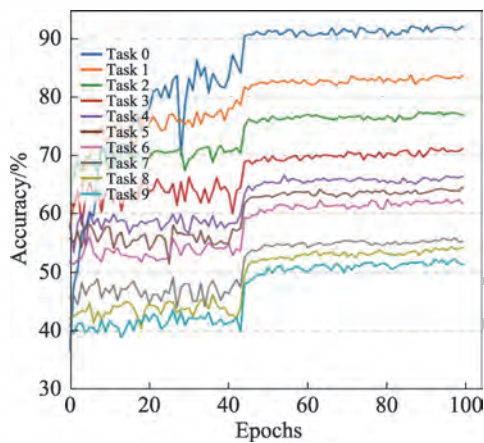


图6 准确率训练曲线

Fig.6 Accuracy training curves

呈现三者收敛过程。本节对上述损失分别进行展示,如图7~图9所示。

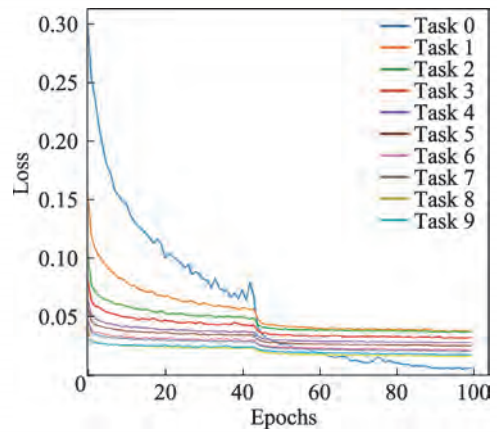


图7 分类损失收敛曲线

Fig.7 Convergence curves of classification loss

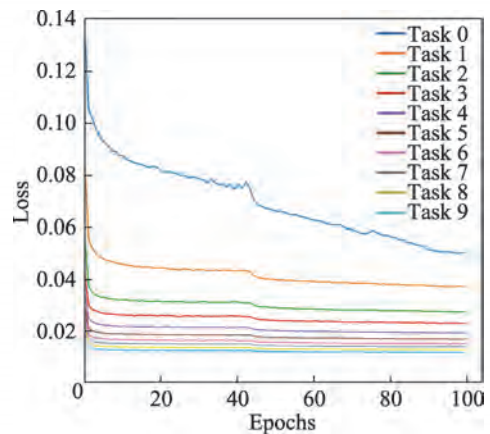


图8 自监督损失收敛曲线

Fig.8 Convergence curves of self-supervised loss

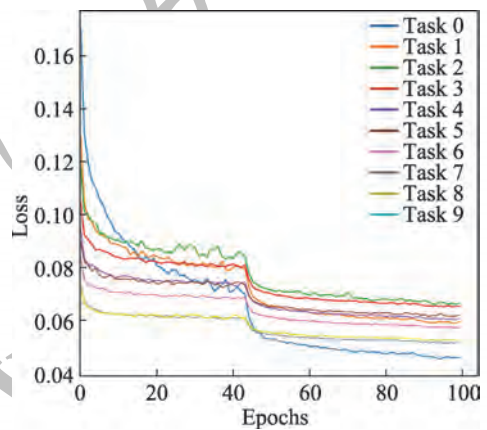


图9 蒸馏损失收敛曲线

Fig.9 Convergence curves of distillation loss

由于基模型的训练不存在旧模型和蒸馏损失,图9中没有Task 0的收敛曲线。如图可见,在各增量任务中随着训练轮次进行,BWSS的各损失函数项具有相对一致的收敛趋势,说明各损失函数能够相互配合,对模型的增量训练是有利的。蒸馏损失的收敛说明模型在增量过程中能够和旧模型保持较为符合的输出期望,即在学习新知识的同时良好地保持了旧知识。这得益于平衡权重和自监督辅助任务缓解了模型的知识遗忘。从整体来看,BWSS具有良好的收敛性且能够保持较为稳定的状态。

### 3.9 复杂度分析

本节分析BWSS与各对比方法的理论复杂度以及模型的参数量(Params)和计算量(FLOPs),结果如表3所示,表中 $n$ 代表输入样例量, $m$ 代表样例内存大小, $t$ 表示模型参数量。

表3 模型复杂度分析  
Table 3 Model complexity analysis

增量模型	复杂度	参数量/ $10^6$	计算量/GFLOPs
LwF	$O(n)$	11.17	0.55
EWC	$O(n)+O(t)$	11.17	0.55
PODNet	$O(n)+O(m)$	11.17	0.55
iCaRL	$O(n)+O(m)$	11.17	0.55
Bic	$O(n)+O(m)$	11.17	0.55
WA	$O(n)+O(m)$	11.17	0.55
BWSS	$O(n)+O(m)$	11.19	0.55

LwF和EWC作为无样例回放的方法,基本复杂度为 $O(n)$ ,其中EWC需要计算模型参数对损失的敏感度因而实际复杂度为 $O(n)+O(t)$ 。PODNet、iCaRL、Bic和WA作为基于样例回放的方法都需要计算样例内存中的样本,因此复杂度为 $O(n)+O(m)$ 。

BWSS主要包含平衡权重和自监督辅助任务两个模块。平衡权重根据各样例自身梯度和平均梯度生成,因此平衡权重算法复杂度为 $O(n)$ 。而自监督辅助任务需要对样例进行4倍旋转扩充操作,因此复杂度为 $O(4n)$ 。又BWSS是基于样例回放的算法,样例回放带来的复杂度为 $O(m)$ ,因此BWSS算法总复杂度为 $O(n)+O(m)$ 。由此可见,BWSS的理论复杂度主要取决于样例量和样例内存的大小,与主流方法一致。

在模型参数量和计算量方面,由于本文所对比的方法都是基于正则约束或偏移修正的类增量学习方法,主要影响因素在于骨干网络。为公平对比,各

对比方法的骨干网络都为Resnet-18,因此计算量和参数量一致。同时,由于BWSS的自监督辅助任务需要引入额外的全连接分类层,模型的参数量对比方法多出 $2 \times 10^4$ 。鉴于BWSS的增量任务准确率优于各对比方法,因此多出的模型参数量是可接受的。

## 4 结束语

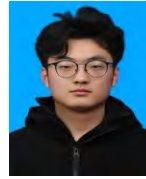
为解决类增量学习中知识灾难性遗忘问题,本文提出一种新的融合平衡权重和自监督的类增量学习方法(BWSS)。通过自监督模块,模型能够挖掘各增量任务数据中的冗余特征和共性特征知识。同时利用旧类样本的预测期望小于新类这一特点,设计了自适应平衡权重来扩大旧样例在同一数据批次中的损失回传占比,有效地发挥旧样例在增量学习中的偏移修正作用。在样本量少的CIFAR-100和样本量多的CIFAR-10数据集,本文模型相比主流类增量学习方法都取得了更好的增量性能表现。通过消融实验和特征可视化,证明了本文方法是有效的。

## 参考文献:

- [1] 刘冰瑶, 刘进锋. 增量学习研究综述[J]. 现代计算机, 2022, 28(13): 72-75.
- [2] LIU B Y, LIU J F. A survey of incremental learning[J]. Modern Computer, 2022, 28(13): 72-75.
- [3] LI Z Z, HOIEM D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935-2947.
- [4] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531, 2015.
- [5] RANNEN A, ALJUNDI R, BLASCHKO M, et al. Encoder based lifelong learning[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Piscataway: IEEE, 2017: 1320-1328.
- [6] KIRKPATRICK J, PASANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521-3526.
- [7] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017: 2994-3003.
- [8] WU C S, HERRANZ L, LIU X, et al. Memory replay GANS: learning to generate new categories without forgetting[C]//Advances in Neural Information Processing Systems 31, Montréal, Dec 3-8, 2018: 5962-5972.



- [8] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. iCaRL: incremental classifier and representation learning[C]//Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Piscataway: IEEE, 2017: 5533-5542.
- [9] HOU S H, PAN X Y, LOY C C, et al. Learning a unified classifier incrementally via rebalancing[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 831-839.
- [10] WU Y, CHEN Y P, WANG L J, et al. Large scale incremental learning[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 374-382.
- [11] ZHAO B W, XIAO X, GAN G J, et al. Maintaining discrimination and fairness in class incremental learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 13205-13214.
- [12] ALJUNDI R, BABILONI F, ELHOSEINY M, et al. Memory aware synapses: learning what (not) to forget[C]//Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 139-154.
- [13] DOUILLARD A, CORD M, OLLION C, et al. PODNet: pooled outputs distillation for small-tasks incremental learning[C]//Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 86-102.
- [14] ZHU F, ZHANG X Y, WANG C, et al. Prototype augmentation and self-supervision for incremental learning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, Jun 19-25, 2021. Piscataway: IEEE, 2021: 5871-5880.
- [15] WANG L X, XU S C, WANG X, et al. Addressing class imbalance in federated learning[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 10165-10173.
- [16] LEE H, HWANG S J, SHIN J W. Self-supervised label augmentation via input transformations[C]//Proceedings of the 37th International Conference on Machine Learning, Jul 13-18, 2020: 5714-5724.
- [17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Piscataway: IEEE, 2016: 770-778.



巩佳义(1998—),男,江苏睢宁人,硕士研究生,主要研究方向为机器学习。

**GONG Jiayi**, born in 1998, M.S. candidate. His research interest is machine learning.



许鑫磊(1997—),男,山东威海人,博士研究生,主要研究方向为机器学习、小样本学习等。

**XU Xinlei**, born in 1997, Ph.D. candidate. His research interests include machine learning, few-shot learning, etc.



肖婷(1990—),女,四川遂宁人,博士,讲师,主要研究方向为机器学习、医学图像处理等。

**XIAO Ting**, born in 1990, Ph.D., lecturer. Her research interests include machine learning, medical image processing, etc.



王喆(1981—),男,江苏淮安人,博士,教授,博士生导师,CCF会员,主要研究方向为模式识别、机器学习等。

**WANG Zhe**, born in 1981, Ph.D., professor, Ph.D. supervisor, CCF member. His research interests include pattern recognition, machine learning, etc.