



基于神经坍塌的类增量学习方法*

扈超舜, 叶标华, 谢晓华, 赖剑煌

中山大学计算机学院, 广东 广州 510006

摘要: 类增量学习中的新旧类不平衡导致少数坍塌发生, 旧类的识别能力降低. 现有方法通常基于经验调整深度特征空间中类别间的几何关系以避免少数坍塌, 缺乏理论指导. 神经坍塌从理论上揭示了类别间的最佳几何结构——等角紧致框架. 受此启发, 本文提出了一种名为持续构造神经坍塌的方法来解决少数坍塌. 该方法通过紧致损失和等角损失来约束形成等角紧致框架结构. 然而不平衡数据分布导致全局质心估计不准确和旧类之间约束困难, 进而导致上述两个损失无法充分施展其效果. 为此, 本文进一步提出了分类器向量辅助模块和难例采样模块来分别解决上述两个问题. 实验结果表明, 本文提出的方法有效诱导了神经坍塌的发生, 并且在 CIFAR100 和 ImageNet 数据集上都超过了当前最优方法.

关键词: 类增量学习; 神经坍塌; 少数坍塌; 动态扩展结构

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 2097-0137(2024)06-0224-12

Inducing Neural Collapse in class-incremental learning

HU Chaoshun, YE Biaohua, XIE Xiaohua, LAI Jianhuang

School of Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Abstract: In class-incremental learning, the imbalance between new and old classes leads to Minority Collapse, resulting in decreased performance for old classes. Existing methods typically rely on empirical adjustments to the geometric relationships between classes in the deep feature space to avoid Minority Collapse, lacking theoretical guidance. Neural Collapse theoretically reveals the optimal geometric structure between classes—the Equiangular Tight Frame (ETF). Inspired by this, this paper proposes a method called Continuous Construction of Neural Collapse (CCNC) to address Minority Collapse. The method constrains the formation of an ETF structure through compactness loss and equiangular loss. The imbalanced data distribution can lead to inaccurate global centroid estimation and difficulties in maintaining constraints among old classes, rendering these losses ineffective. To address the above two issues, the paper presents a classifier vector supplementation module and a hard example sampling module, respectively. Experimental results indicate that the proposed method successfully induces Neural Collapse and outperforms the current best methods on the CIFAR100 and ImageNet datasets.

Key words: class-incremental learning; Neural Collapse; Minority Collapse; dynamically expanding architecture

* 收稿日期: 2024-04-29 录用日期: 2024-05-17 网络首发日期: 2024-07-31

基金项目: 国家自然科学基金(U22A2095, 62076258)

作者简介: 扈超舜(1997年生), 男; 研究方向: 模式识别、计算机视觉;

E-mail: huchsh3@mail2.sysu.edu.cn

通信作者: 赖剑煌(1964年生), 男; 研究方向: 模式识别、计算机视觉;

E-mail: stsljh@mail.sysu.edu.cn

全文阅读



ZR20240136

类增量学习旨在不断学习新类别以适应不断变化的环境. 因隐私保护或训练成本, 模型通常仅能保存少量或无法保存任何旧样本. 新旧类别之间的样本数量不平衡导致训练时旧类的梯度过小, 进而出现少数坍塌(MC, Minority Collapse)的问题(Fang et al., 2021). 少数坍塌表明, 在学习新知识过程中不同的旧类质心逐渐靠近导致旧类间的分类界限模糊, 降低了旧类的识别能力.

为了调节不同类之间的几何关系(位置、角度等等)以避免少数坍塌, 现有的方法大致可以分为三类. 一类方法提出了多种筛选算法保留关键样本以强化类别边界(Rebuffi et al., 2017; Chaudhry et al., 2018; Aljundi et al., 2019; Bang et al., 2021; Zhao et al., 2022; Luo et al., 2023). 另一类方法则强调限制参数的变动来提高深度特征空间的稳定性(Kirkpatrick et al., 2017; Zenke et al., 2017; Lee et al., 2020; Yan et al., 2021; Wang et al., 2022; Cai et al., 2023; Huang et al., 2023; Wang et al., 2023; Yang et al., 2023a), 进而防止旧类质心靠近. 还有一类方法通过知识蒸馏约束新旧空间中类别的几何关系保持一致(Li et al., 2018; Dhar et al., 2019; Wu et al., 2019; Douillard et al., 2020; Tao et al., 2020; Gao et al., 2022; Kang et al., 2022).

然而, 上述工作大多依据实践经验来调整类间的几何关系, 缺乏坚实的理论支撑. 近期, 神经坍塌(NC, Neural Collapse)从理论分析和实验证明揭示了不同类在特征向量空间和分类器向量空间中的最佳几何关系(Papayan et al., 2020), 即紧致和等角. “紧致”意味着同一类别的特征向量及其分类器向量坍塌到同一顶点, “等角”意味着任何两个顶点之间的角度相同且最大化. 此结构数学上归类为等角紧致框架(ETF, Equiangular Tight Frame), 使得费雪判别比(Fisher, 1936)最大化, 表现出卓越的泛化能力和鲁棒性(E et al., 2020; Han et al., 2021; Ji et al., 2021; Lu et al., 2022; Mixon et al., 2022; Zhou et al., 2022).

为缓解少数坍塌的问题, 我们提出了一种名为“持续构造神经坍塌”(CCNC, Continuous Construction of Neural Collapse)的方法. 首先, 我们提出了紧致损失和等角损失来促进神经坍塌的形成. 其次, 为了解决不平衡数据分布导致损失无法充分施展其效果的问题, 我们提出了分类器向量辅助模块和难例采样模块. 分类器向量辅助模块将所有的分类器向量视为对应类的质心, 并且在每个训练批次中参与约束. 通过这种方式, 更有效地估计了全局质心, 从而更准确地估计了不同类之间的角度. 第二个模块是难例采样模块. 该模块通过方差指标识别等角性较差的旧类别, 并增加这些类别的样本采样概率, 从而在每个训练批次中额外采样部分难例旧样本更好约束旧类之间的角度. 我们的方法有效地诱导了神经坍塌的发生, 并在CIFAR100和ImageNet数据集上都超越了当前最优的方法.

1 相关工作

类增量学习 一些工作尝试筛选部分关键旧样本来明确类间边界. iCaRL(Rebuffi et al., 2017)采用了贪婪策略保留最靠近类别质心的样本, 从而在交叉熵优化过程中使得旧质心远离新质心. RWalk(Chaudhry et al., 2018)和Rainbow(Bang et al., 2021)则倾向于保留边界样本, 旨在清晰界定不同类别的界限, 确保类间具有明显的间隔. GSS(Aljundi et al., 2019)则从梯度上筛选出尽量少破坏旧类之间角度的新样本. Zhao et al.(2022)和Luo et al.(2023)提出了下采样策略来在同样的存储空间下保留更多的旧样本. 还有一些工作致力于减小重要参数的变动从而保持深度特征空间的稳定性. EWC(Kirkpatrick et al., 2017)通过费歇尔信息矩阵的逆对角元素来反映参数的重要性, 并用于指导参数更新时的稳定性约束. 后续的Zenke et al.(2017)、Lee et al.(2020)和Yang et al.(2023a)尝试寻找更有效的先验知识来计算参数重要性. 然而, 依赖先验知识评估参数重要性存在局限, 主要因为没有一种通用的先验适用于所有数据集, 且模型无法预见未来的数据分布, 使得参数重要性评估变得复杂且困难. 为此, Yan et al.(2021)提出了动态扩展结构(DEA, Dynamic Expansion Architectures)的概念, 直接冻结旧特征提取器以保持旧特征, 并添加新特征提取器学习新类别的特征. 这种结构因其优异性能, 被后续研究广泛采纳和改良(Wang et al., 2022; Cai et al., 2023; Huang et al., 2023; Wang et al., 2023). 还有一些工作通过知识蒸馏约束新旧空间中不同旧类间的几何关系保持一致. Li et al.(2018)和Wu et al.(2019)通过引入对数知识蒸馏(Hinton et al., 2015)来间接约束不同的类在新旧空间中的几何关系保持一致. Dhar et al.(2019)、Douillard et al.(2020)和Kang et al.(2022)则通过特征蒸馏

(Romero et al., 2014)约束新旧模型在不同层级的特征上保持一致. Tao et al.(2020)和 Gao et al.(2022)则通过关系蒸馏(Park et al., 2019)来直接约束新空间中不同旧类之间的几何关系尽量和旧空间中的保持一致.

神经坍塌 Papyan et al.(2020)首先发现在交叉熵损失的优化下神经坍塌现象会自然出现. 下面首先正式地定义神经坍塌. 假设总共有 C 个类别, \mathcal{G} 为特征提取器, 第 c 个类别的所有样本的集合表示为 \mathcal{D}_c , 第 c 个类别对应的分类器向量表示为 w_c , μ_c 表示 \mathcal{D}_c 中所有特征的均值, μ_G 表示所有特征的均值, 即 $\mu_G := \frac{1}{C} \sum_{c=1}^C \mu_c$. 其可以表示为以下 4 个现象:

1) (NC1)类内变异性坍塌 对任意的 $x \in \mathcal{D}_c$ 都有 $\mathcal{G}(x) = \mu_c$. 也即同一个类的所有样本坍塌到深度特征空间上的同一个顶点.

2) (NC2)收敛于等角紧致框架 对所有 c, c' 均有

$$\|\mu_c - \mu_{c'}\|_2 = \|\mu_{c'} - \mu_c\|_2.$$

设 $\tilde{\mu}_c := \frac{\mu_c - \mu_G}{\|\mu_c - \mu_G\|_2}$. 指示器变量 $\delta_{\{\text{条件}\}}$ 判断条件是否成立? 若条件成立则输出为 1, 否则输出为 0. 例如:

$$\delta_{\{i=j\}} = \begin{cases} 1, & \text{当条件 } i=j \text{ 成立,} \\ 0, & \text{当条件 } i \neq j \text{ 不成立.} \end{cases}$$

$\langle \cdot, \cdot \rangle$ 表示两个向量的内积. 则

$$\langle \tilde{\mu}_c, \tilde{\mu}_{c'} \rangle = \frac{C}{C-1} \delta_{\{c=c'\}} - \frac{1}{C-1},$$

也即不同的顶点之间的角度保持一致且余弦值为 $-\frac{1}{C-1}$.

3) (NC3)自对偶性 对所有的 c , 都存在:

$$\frac{w_c}{\|w_c\|_F} = \frac{\tilde{\mu}_c}{\|\tilde{\mu}_c\|_F},$$

也即分类器向量也会坍塌到顶点.

4) (NC4)简化为最近邻(NCC, Nearest Class-Center)分类器 即对于任意样本, 分类器概率最高的类等价于 NCC 分类器中距离最近的质心对应的类.

随后许多学者发现, 无论使用何种损失函数训练模型, 这一现象普遍存在. 后续的理论推导和实验 (E et al., 2020; Han et al., 2021; Ji et al., 2021; Lu et al., 2022; Mixon et al., 2022; Zhou et al., 2022) 表明, 这种结构是全局最优解, 表现出优异的泛化能力和鲁棒性. Fang et al.(2021)发现在不平衡数据集上神经坍塌并未发生. 反而不同旧类别之间不能保持理想的角度, 近似坍塌到同一个顶点. 为了解决这个问题, 一些研究引入了 ETF 分类器来替代可学习的分类器 (Yang et al., 2022; Li et al., 2023; Yang et al., 2023b; Fu et al., 2024). 这是一组预先分配的、不可训练的分类器向量, 并且满足 ETF 结构. 然而, 此类分类器并不适用于需要持续添加新类别的增量学习场景.

2 方 法

设一个训练批次的样本数量为 b , \mathbf{A} 为余弦值矩阵, A_{ij} 为一个训练批次中第 i 个样本和第 j 个样本之间的角度的余弦值 (样本均 L2 范数归一化), C 为类别总数, c_i 表示第 i 个样本的类别序号, S 表示同一个训练批次中所有不同类别的样本对的数量. 一种直观的思想是直接构造紧致损失约束同一个类的样本坍塌到同一个顶点 (余弦值为 1):

$$\mathcal{L}_{CL} = \frac{1}{b^2 - S} \sum_{i=1}^b \sum_{j=1}^b \delta_{\{c_i=c_j\}} \|A_{ij}, 1\|_{\text{SmoothL1}},$$

这里 $\|\cdot\|_{\text{SmoothL1}}$ 表示计算 SmoothL1 距离. 同时构造等角损失来约束不同类之间样本的角度的余弦值保持为 $-\frac{1}{C-1}$:

$$\mathcal{L}_{EL} = \frac{1}{S} \sum_{i=1}^b \sum_{j=1}^b \delta_{\{c_i \neq c_j\}} \left\| A_{ij}, -\frac{1}{C-1} \right\|_{\text{SmoothL1}}.$$

然而不平衡数据导致两个损失不能充分发挥其作用, 具体表现为两个关键的障碍. 其一, 角度的计算需要减去全局质心 (所有样本的均值). 由于每个训练批次中旧类样本的稀缺, 直接计算出的质心往往偏向于新类. 最终导致角度计算不够准确. 其次, 每个批次中旧类类别和数量过少使得很难约束旧类之间的角度, 主要约束的常常是新类之间和旧类之间的角度. 本文首先从梯度的角度发现普遍采用的平衡交叉熵损失并不能很好解决上述问题, 然后分别提出了分类器向量辅助模块和难例采样模块来解决上述问题.

2.1 梯度分析

少数坍塌发生的根本原因是旧类样本过少导致的梯度过小 (Fang et al., 2021). 虽然在类增量学习领域常用的平衡交叉熵损失 (Ren et al., 2020) 等价于增大旧类的梯度, 但依然无法使得神经坍塌自然发生. 接下来我们从梯度的角度来说明以上结论. 由于每次更新梯度都是按照一个训练批次来更新, 因此接下来重点考虑一个批次内的梯度. 假设一个批次含有 b 个样本, 则第 i 个样本的特征可以表示为 $\phi_i (i = 1, 2, \dots, b)$, 其对应的类别的序号为 $c_i (i = 1, 2, \dots, b)$. 第 j 个分类器向量可以表示为 $w_j (j = 1, 2, \dots, C)$. 则第 i 个样本的交叉熵损失可以表示为:

$$\mathcal{L}_{CE} = - \sum_{n=1}^C \delta_{\{n=c_i\}} \log(\hat{y}_n),$$

其中 $\hat{y}_n = \frac{e^{w_n \cdot \phi_i}}{\sum_{m=1}^C e^{w_m \cdot \phi_i}}$. 许多方法提出了多种平衡交叉熵损失, 但是这些交叉熵损失可以看作是简单修改 \hat{y}_j 来增

大旧类的梯度, 减少新类的梯度. \hat{y}_j 的调整不会影响我们的分析, 为了简洁起见, 我们接下来依然对交叉熵损失进行梯度分析, 然后讨论平衡交叉熵的影响. 则对于第 i 个样本而言, 第 j 个分类器向量 w_j 的梯度可以表示为:

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \begin{cases} (\hat{y}_j - 1) \cdot \phi_i & \text{if } j = c_i, \\ \hat{y}_j \cdot \phi_i & \text{if } j \neq c_i. \end{cases} \quad (1)$$

从上述梯度分析中可以看到, 对于分类器向量 w_j 而言, 如果输入样本的类别序号为 j , 则损失会拉近特征 ϕ_i 和 w_j 的距离, 否则会推远 ϕ_i 和 w_j 的距离. 接下来统计一个训练批次下 w_j 的梯度, 假设旧类的数量为 C_o , 则有:

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \underbrace{\sum_{i=1}^b \delta_{\{j=c_i\}} (\hat{y}_j - 1) \cdot \phi_i}_{\text{拉近同类}} + \underbrace{\sum_{i=1}^b \delta_{\{j \neq c_i \wedge c_i \leq C_o\}} \hat{y}_j \cdot \phi_i}_{\text{推远旧类}} + \underbrace{\sum_{i=1}^b \delta_{\{j \neq c_i \wedge c_i > C_o\}} \hat{y}_j \cdot \phi_i}_{\text{推远新类}}.$$

类似于公式(1), 首先将其分为拉近和推远, 然后更进一步, 将推远细分为推远旧类和推远新类. 在不平衡数据中, 每个批次中旧类样本非常少, 对于旧类的 w_j 而言, 第三项远离新类的梯度最大, 这使得新旧类之间的夹角过大. 并且由于该项的梯度占据主导, 因此不同的旧类的 w_j 的梯度方向基本一致, 这导致了旧类之间互相靠近, 进而导致了少数坍塌. 平衡交叉熵损失通过减小旧类的 \hat{y}_j 的从而增大旧类的梯度, 通过增大新类的 \hat{y}_j 的从而减小新类的梯度. 对应到公式中就是第一项和第二项会增大, 第三项会变小, 有效克制了新旧类之间的夹角过大的问题. 然而在一个批次中旧类的类别数量过少, 例如在 ImageNet1000-B0 的评估协议下的最后一个阶段, 每个批次设置的样本数量为 256, 虽然平均有 34.13 个旧类样本, 但是旧类别的数量高达 900. 因此旧类之间难以很好地远离以形成神经坍塌.

2.2 分类器向量辅助模块

神经坍塌的现象 (NC3) 表明了分类器向量最终也会坍塌到对应顶点. 受此启发, 本文直接将所有的分类器向量作为对应类的质心纳入到每个批次中参与角度约束. 每个训练批次中每个类别都至少有一个“样本”参与计算全局质心, 从而更准确地计算不同样本之间的角度 (详见算法 1).

在算法 1 中, 本文详细描述了分类器向量辅助模块如何与紧致损失和等角损失结合. 假设每个批次共

算法 1 结合分类器向量补充模块的等角和紧凑损失

```

输入: 特征  $\Phi$ ; 标签  $\mathbf{y}$ ; 分类器向量  $\mathbf{W}$ .
输出: 结合分类器向量补充模块的紧凑损失  $\mathcal{L}_{CCL}$ ; 结合分类器向量补充模块的等角损失  $\mathcal{L}_{CEL}$ .
1:  $\Phi' \leftarrow [\Phi, \mathbf{W}]$ 
2:  $\mathbf{y}' \leftarrow [\mathbf{y}, [0, 1, \dots, C-1]]$ 
3:  $\mathbf{cs} \leftarrow \mathbf{0}$ 
4: for  $i \in \{1, 2, \dots, C\}$  do
5:    $\mathbf{cs} \leftarrow \mathbf{cs} + \text{Mean}(\Phi'[\mathbf{y}' == i])$ 
6: end for
7:  $\mathbf{gc} \leftarrow \mathbf{cs}/C$ 
8:  $\Phi' \leftarrow \Phi' - \mathbf{gc}$ 
9:  $\Phi' \leftarrow \Phi' / \text{StopGrad}(|\Phi'|_2)$ 
10:  $\mathbf{A} = \Phi' \cdot \Phi'^T$ 
11:  $\mathbf{Y} \leftarrow \text{OnehotMatrix}(\mathbf{y}')$ 
12:  $\mathbf{M} = \mathbf{Y} \cdot \mathbf{Y}^T$ 
13:  $\mathbf{Y}[\mathbf{C}_0 :, \mathbf{C}_0 :] \leftarrow \mathbf{0}$ 
14:  $\mathbf{M}_0 = \mathbf{Y} \cdot \mathbf{Y}^T$ 
15: 基于公式(2)计算  $\mathcal{L}_{CCL}$ 
16: 基于公式(3)计算  $\mathcal{L}_{CEL}$ 

```

有 b 个样本, 这些样本的特征构成了矩阵 Φ , 对应的类别序号构成了向量 \mathbf{y} . C 为全部类别的数量, \mathbf{C}_0 是旧类别的数量, $[a, b]$ 表示将集合 a 和集合 b 中的元素合并为一个新的集合. 分类器向量矩阵 $\mathbf{W}([\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_c])$ 被视为类别质心的集合, 并与所有特征一起参与计算(第 1~2 行):

$$\begin{aligned}\Phi' &\leftarrow [\Phi, \mathbf{W}], \\ \mathbf{y}' &\leftarrow [\mathbf{y}, (1, 2, \dots, C)].\end{aligned}$$

此时所有的特征总计 $b + C$ 个. 值得注意的是, 当前批次中的每个类别至少有一个特征向量. 接着通过求均值来分别得到每个类的质心, 然后求和得到 \mathbf{cs} , 然后对 \mathbf{cs} 求平均以得到全局质心 \mathbf{gc} (第 3~7 行):

$$\mathbf{gc} \leftarrow \frac{\mathbf{cs}}{C}.$$

通过这种方式, 同时解决了两个导致全局质心估计出现偏差问题. 一是旧类类别缺失的问题, 二是旧类的样本数量偏少的问题. 然后通过减去 \mathbf{gc} 来中心化(第 8 行). 随后根据神经坍塌的要求, 进行 L2 范数归一化确保得到的余弦值在 $[-1, 1]$ 的范围内(第 9 行), 其中 StopGrad 表示不进行反向传播:

$$\Phi' \leftarrow \Phi' / \text{StopGrad}(|\Phi'|_2).$$

最后通过将 Φ' 与其转置相乘来计算不同样本之间的角度的余弦值(第 10 行), 得到余弦值矩阵 \mathbf{A} . 接下来构造掩码 \mathbf{M} , 一个 $(b + C) \times (b + C)$ 矩阵. 第 i 行第 j 列为 1 表示第 i 个样本和第 j 个样本的类别相同, 0 则表示类别不同(第 11~12 行). 掩码 \mathbf{M}_0 与 \mathbf{M} 的区别在于只有旧类的样本类别相同时则为 1, 其他均为 0(第 13~14 行). 基于 \mathbf{M}_0 , 使用 SmoothL1 损失鼓励同一个旧类的样本和分类器向量坍塌到同一顶点. 紧致损失 \mathcal{L}_{CCL} (第 15 行)可以表示为:

$$\mathcal{L}_{CCL} = \text{Mean}(\| \mathbf{A}, \mathbf{1} \|_{\text{SmoothL1}} \cdot \mathbf{M}_0), \quad (2)$$

这里 $\| \cdot \|_{\text{SmoothL1}}$ 表示对矩阵中的每一个元素计算 SmoothL1 距离. 余弦值为 1 表示两向量 (经过 L2 范数归一化后的向量) 坍塌到同一顶点. 基于 \mathbf{M} , 使用 SmoothL1 损失将类间角度收敛至等大. 等角损失 \mathcal{L}_{CEL} (第 16 行) 可以表示为:

$$\mathcal{L}_{CEL} = \text{Mean} \left(\left\| \mathbf{A}, -\frac{1}{(C-1)} \right\|_{\text{SmoothL1}} \cdot (\mathbf{1} - \mathbf{M}) \right), \quad (3)$$

$-\frac{1}{(C-1)}$ 是神经坍塌揭示的理论上的最佳角度的余弦值.

2.3 难例采样模块

对于旧类而言, 即使引入了分类器向量, 旧类的特征 ϕ 依然以很低的频次参与优化, 不同旧类样本在单个训练批次中相遇的机会有限, 导致难以有效约束旧类间的几何关系. 鉴于此, 本文提出难例采样模块在每个训练批次中额外选取一小部分旧样本参与约束. 我们认为应赋予难例样本更高的采样权重, 故提出了一个难例优先的采样概率计算方法. 首先构建一个队列 \mathcal{Q} 保存每个类最新提取的 n 个特征. 每个训练批次结束后更新该队列. 接着依据神经坍塌理论, 对队列 \mathcal{Q} 中的所有特征进行中心化和 L2 范数归一化, 最后计算不同类质心之间角度的余弦值, 得到余弦值矩阵 \mathbf{A} . A_{ij} 表示第 i 个类质心和第 j 个类质心之间角度的余弦值. 接下来计算每个类质心和其他类质心之间角度的余弦值和最佳角度的余弦值 $-\frac{1}{C-1}$ 之间的方差:

$$\sigma_i^2 = \frac{1}{C-1} \sum_{j=1, j \neq i}^C (A_{ij} + 1/(C-1))^2.$$

该方差衡量了不同类满足等角性的程度, 方差越大则说明越不满足等角性, 越应该优先考虑. 对于第 k 个样本的采样概率 p_k , 对应类的方差越大则采样概率更高, 假设 c_k 为第 k 个样本对应类的序号, 则

$$p_k = \frac{\sigma_{c_k}^2}{\sum_{l=1}^C \sigma_{c_l}^2}.$$

每次选中的样本之后不再参与选择直到所有的样本都被选择了一次之后则重置.

2.4 损失优化

我们结合动态扩展结构 (Yan et al., 2021) 来更好地减少灾难性遗忘的同时保护旧的 ETF 结构. 以下概述了整个持续构造神经坍塌的过程. 假设总共 T 个阶段, 第 t 个阶段的数据集为 D_t . 第一阶段, 特征提取器 \mathcal{G}_1 和分类器 \mathcal{F}_1 在交叉熵损失 \mathcal{L}_{CE} 、结合分类器向量辅助模块的紧凑损失 \mathcal{L}_{CCL} 和等角损失 \mathcal{L}_{CEL} 的优化下在数据集 D_1 上进行训练:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{CEL} + \beta \cdot \mathcal{L}_{CCL}, \quad (4)$$

其中 α 和 β 是超参数. 在随后的第 t 个阶段 $t \in \{2, 3, \dots, T\}$, 冻结所有特征提取器以保护旧类间的等角性. 然后引入一个新的特征提取器 \mathcal{G}_t 、一个辅助分类器 \mathcal{F}_t^a 和一个新的分类器 \mathcal{F}_t . 辅助分类器只会在当前阶段使用, 并不会传递到下一个阶段. 旧分类器 \mathcal{F}_{t-1} 的参数矩阵 \mathbf{W} 被赋值给新分类器 \mathcal{F}_t 的参数矩阵的对应部分. 被赋值的部分不再更新以保护分类器向量之间的等角性. 样本通过所有特征提取器然后在维度上进行拼接以得到特征 ϕ . 每个阶段的学习过程又分为两个步骤: 第一个步骤主要聚焦于新类别特征的学习, 其中新特征提取器提取的特征被送入 \mathcal{F}_t^a 并通过交叉熵损失进行优化:

$$\mathcal{L}_{\text{step1}} = \mathcal{L}_{CE}.$$

第二个步骤旨在诱导神经坍塌的发生. 设原始采样的训练批次设为 B , 经过难例采样模块采样的训练批次设为 B_k . 对于 B , ϕ 被输入 \mathcal{F}_t 并使用平衡交叉熵损失 (Ren et al., 2020) 优化. 此外, 使用结合分类器向量辅助模块的 \mathcal{L}_{CEL} 和 \mathcal{L}_{CCL} 来促进神经坍塌的发生. 对于 B_k , 只使用结合分类器向量辅助模块的 \mathcal{L}_{CEL} 和 \mathcal{L}_{CCL} 来促进神经坍塌的发生. 这里为了简单起见, 设置两个批次中的 \mathcal{L}_{CEL} 的系数相同, \mathcal{L}_{CCL} 的系数也相同.

于是同一个类型的损失可以合并简写为以下形式:

$$\mathcal{L}_{step2} = \mathcal{L}_{BCE} + \alpha \cdot \mathcal{L}_{CEL} + \beta \cdot \mathcal{L}_{CCL},$$

其中 α 和 β 是与式(4)相同的超参数. 尽管存在两个步骤, 但总的训练轮次与其他一步骤的方法保持一致.

3 实验

3.1 实验设定和实施细则

数据集 在广泛使用的数据集 CIFAR100(Krizhevsky, 2009)和 ImageNet100/1000(Deng et al., 2009)进行了评估. CIFAR100 包含 100 个类别, 每个类别有 500 张训练图像和 100 张评估图像, 图像分辨率为 32×32 . ImageNet1000 是一个大规模数据集, 包含 1 000 个类别, 总计 128 万张训练图像. ImageNet100 是从 ImageNet1000 中随机选取 100 个类别构成的数据集. 我们所选的 100 个类别与 DER(Yan et al., 2021)保持一致.

评估协议 遵循 iCaRL(Rebuffi et al., 2017)的设置, 我们采用了 5 种广泛使用的评估协议: 1) CIFAR100-B0: 将 CIFAR100 的 100 个类别均等地分为 5 个、10 个和 20 个任务, 可保留样本不超过 2 000 个. 2) CIFAR100-B50: 先将 CIFAR100 中的 50 个类别进行预训练, 然后将剩余的 50 个类别均等地分为 5 个和 10 个任务, 每类保留 20 个样本. 3) ImageNet100-B0: 将 ImageNet100 的 100 个类别均等地分为 10 个任务, 可保留样本同样不超过 2 000 个. 4) ImageNet100-B50: 先将 ImageNet100 中的 50 个类别进行预训练, 然后将剩下的 50 个类别均分为 10 个任务, 每类保留 20 个样本. 5) ImageNet1000-B0: 将 ImageNet1000 的 1 000 个类别均等地分为 10 个任务, 可保留样本不超过 20 000 个.

实施细则 对于 CIFAR100 上的实验, 我们采用了一个调整过的 ResNet-18(Yan et al., 2021)作为特征提取器, 并将批次大小设置为 128. 对于 ImageNet 上的实验, 我们使用标准的 ResNet-18(He et al., 2016)作为特征提取器, 批次大小设为 256. 下面是在两个数据集上相同的参数设置. 第一阶段, 我们将初始学习率设定为 0.1, 并利用余弦退火策略逐渐衰减学习率至 0. 训练轮次设置为 200 轮. 采用动量为 0.9、权重衰减为 $5e$ 的随机梯度下降(SGD)优化器. 后续的阶段中, 每个阶段分为两个步骤. 第一个步骤, 学习率设置为 0.1, 并利用余弦退火策略逐渐衰减学习率至 0, 训练轮次设置为 170 轮. 采用动量为 0.9、权重衰减为 $5e$ 的随机梯度下降(SGD)优化器. 第二个步骤, 除了将初始学习率调整为 0.001 外, 其他所有学习率相关设置保持不变. 训练轮次为 30 轮. 等角损失和紧致损失的系数分别设置为 2 和 1. 难例采样模块在每个训练批次中采样 32 个难例旧样本. 遵循 Wang et al.(2022)、Cai et al.(2023)、Chen et al.(2023)和 Huang et al.(2023)的设定, 我们采用的数据增强包括自动增强(Cubuk et al., 2019)、随机裁剪、水平翻转以及标准化.

3.2 实验结果与分析

在 CIFAR100 上的实验结果 本文首先在 CIFAR100 数据集上对比相关的类增量学习方法, 包含 iCaRL(Rebuffi et al., 2017), UCIR(Hou et al., 2019), BiC(Wu et al., 2019), WA(Zhao et al., 2020), PODNet(Douillard et al., 2020), DyTox(Douillard et al., 2022), FOSTER(Wang et al., 2022)和 MAFDRC(Chen et al., 2023)等非动态扩展结构方法, 以及 DER(Yan et al., 2021), MCTD(Cai et al., 2023)和 TCIL(Huang et al., 2023)等动态扩展结构的方法, 结果如表 1 所示. 参数量表示训练完所有类别后的最终参数量. 平均准确率表示所有阶段结束时的准确率的平均值. 从准确率上来看, CCNC(本文提出的方法)在 CIFAR100 数据集上都超越了最佳结果. 在 CIFAR100-B0S5、CIFAR100-B0S10、CIFAR100-B0S20、CIFAR100-B50S5 和 CIFAR100-B50S10 上, 结果分别超过最佳方法 **0.68%**、**2.42%**、**1.82%**、**0.62%** 和 **0.25%**. 尤其值得注意的是, 我们的方法在 CIFAR100-B0S10、CIFAR100-B0S20 评估协议下大幅度领先已有的最佳算法.

在 ImageNet 上的实验结果 然后本文在 ImageNet 数据集上对比相关的类增量学习方法, 包括 UCIR(Hou et al., 2019), PODNet(Douillard et al., 2020), TPCIL(Tao et al., 2020), FOSTER(Wang et al., 2022), MAFDRC(Chen et al., 2023), DER(Yan et al., 2021), MCTD(Cai et al., 2023), TCIL(Huang et al., 2023)和 BEEF-C(Wang et al., 2023), 结果如表 2 所示. 从准确率上来看, CCNC(本文提出的方法)在所有的评估协议下均优于现有方法. 具体来说, 在 ImageNet100-B0、ImageNet100-B50 以及 ImageNet1000-B0 的评估协议下, 平均准确率分别超过了现有最佳方法 0.59%、1.27% 和 1.88%.

表1 在CIFAR100数据集上的实验结果
Table 1 Test results on CIFAR100

方法	发表会议	CIFAR100-B0						CIFAR100-B50			
		5个任务		10个任务		20个任务		5个任务		10个任务	
		参数量	平均准确率/%	参数量	平均准确率/%	参数量	平均准确率/%	参数量	平均准确率/%	参数量	平均准确率/%
iCaRL	CVPR'17	11.2	71.14	11.2	61.20	11.2	61.20	11.2	65.06	11.2	71.14
UCIR	CVPR'19	11.2	62.77	11.2	58.17	11.2	58.17	11.2	64.28	11.2	62.77
BiC	CVPR'19	11.2	73.10	11.2	66.48	11.2	66.48	11.2	66.62	11.2	73.10
WA	CVPR'20	11.2	72.81	11.2	67.33	11.2	67.33	11.2	64.01	11.2	72.81
PODNet	ECCV'20	11.2	66.70	11.2	53.97	11.2	53.97	11.2	67.25	11.2	66.70
DyTox	CVPR'22	10.7	73.66	10.7	67.30	10.7	67.30	-	-	-	-
FOSTER	ECCV'22	11.2	77.61	11.2	75.18	11.2	72.26	11.2	75.11	11.2	70.21
MAFDRC	ICCV'23	11.2	78.70	11.2	76.93	11.2	74.09	11.2	74.95	11.2	72.26
DER	CVPR'21	56.0	79.03	112	78.13	224	77.85	67.2	77.15	123.2	75.58
MCTD	CVPR'23	67.2	78.15	123.2	77.40	235.2	76.20	78.4	76.19	134.4	75.43
TCIL	AAAI'23	56.0	80.23	112	79.12	224	78.10	67.2	77.76	123.2	76.91
CCNC	本文方法	56.0	80.92	112	81.54	224	79.92	67.2	78.39	123.2	77.16

“-”表示该方法没有在对应评估协议下测量。

表2 在ImageNet数据集上的对比结果
Table 2 Test results on ImageNet

方法	发表会议	ImageNet100-B50		ImageNet100-B0		ImageNet1000-B0	
		参数量	平均准确率/%	参数量	平均准确率/%	参数量	平均准确率/%
UCIR	CVPR'19	11.2	68.09	-	-	-	-
PODNet	ECCV'20	11.2	74.33	-	-	-	-
TPCIL	ECCV'20	11.2	74.81	-	-	-	-
FOSTER	ECCV'22	11.2	77.54	11.2	78.40	11.2	68.34
MAFDRC	ICCV'23	11.2	77.95	11.2	79.66	11.2	69.37
DER	CVPR'21	123.2	77.74	112	79.81	112	69.81
MCTD	CVPR'23	134.4	79.83	123.2	80.46	123.2	70.08
TCIL	AAAI'23	-	-	112	77.66	-	-
BEEF-C	ICLR'23	-	-	112	79.34	-	-
CCNC	本文方法	123.2	80.42	112	81.73	112	71.96

“-”表示该方法没有在对应评估协议下测量。

3.3 消融实验与分析

消融实验 本文首先对本文提出的各个组件进行消融实验, 结果如表3所示. 本文提出的模型主要由四部分组成: 基线(DER)、等角损失和紧致损失、分类器向量辅助模块与难例采样模块. 在表格中, 它们分别表示为基线、损失、分类器辅助和难例采样. 平均准确率一列表示所有阶段结束时的准确率的平均值. 最终准确率一列表示在模型学习了所有任务之后在所有类别上的总体准确率. 从第一行和第二行的对比结果可以看出, 使用等角损失和紧致损失可以带来最终准确率上1.54%的提升. 接下来的分类器向量辅助模块则可以在此基础上带来1.89%左右的提升. 难例采样模块则继续带来0.7%左右的提升.

表 3 各个组件的消融实验
Table 3 Ablation study results for various components

基线	损失	分类器辅助	难例采样	平均准确率/%	最终准确率/%
√				78.13	68.82
√	√			79.41	70.36
√	√	√		80.80	72.25
√	√	√	√	81.54	72.95

“√”表示添加该模块；空白表示不添加该模块。

灵敏度分析 接下来本文对等角损失的系数 α 和紧致损失的系数 β 进行灵敏度分析,结果如图1所示.横坐标代表取值,纵坐标代表平均准确率.在CIFAR100-B0S10评估协议上,在 $[0, 3]$ 区间内以0.5的增量进行测试.对于 α 而言,在 $[0, 3]$ 范围内,准确率呈现出先上升后下降的趋势.并且即使在比较大的系数下,准确率下降的幅度相对较小,具有良好的稳定性.最佳取值为2.对于 β 而言,在 $[0, 3]$ 范围内,准确率同样呈现出先上升后下降的趋势.增大系数后的下降幅度也相对比较平缓.最佳取值为1.

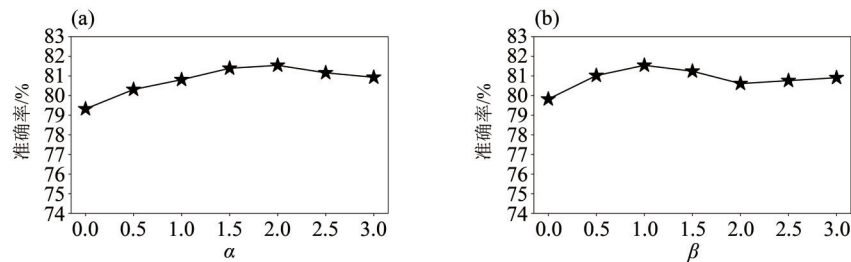


图 1 等角损失(a)和紧致损失(b)的系数的灵敏度分析

Fig. 1 Sensitivity analysis of LCEL (a) and LCCL (b)

等角分析 接下来本文验证提出的方法是否有助于不同的类别保持等角性.具体而言,计算两两不同类别特征向量之间的角度然后计算所有角度的方差,方差越小则说明越趋近于等角.同时也计算了两两分类器向量之间的等角性,结果如图2所示.

本文分别选择了5个模型进行对比.这些方法分别是使用完整数据集训练出的模型(记为“Raw”)、非动态扩展结构的方法WA(Zhao et al., 2020)和BiC(Wu et al., 2019),以及动态扩展结构方法DER(Yan et al., 2021)和本文提出的方法(记为“Ours”).横坐标代表增量的阶段.纵坐标代表方差.(a)和(c)分别展示了在分类器向量空间和特征向量空间中的等角性.可以看到,DER的等角性甚至不如非动态扩展结构的方法.而本文提出的方法十分有效地保持了不同顶点之间的等角性.(b)和(d)则分别展示了在分类器向量空间和特征向量空间中的旧类的等角性.其中虚线表示旧类在学习新知识前(更新前)的等角性,实线表示在学习新知识后(更新后)的等角性,可以看到,由于DER没有约束等角性,因此旧类的等角性在不断地变差.并且由于冻结旧空间使得无法修改旧有的错误信息,从而造成累计误差.而本文提出的方法在学习新知识后几乎不会使得旧类的等角性变差.

4 结 论

在本文中,我们提出了一种名为持续构造神经坍塌的方法在理论指导下解决类增量学习中的少数坍塌问题,即旧类间过于靠近导致难以区分.神经坍塌理论揭示了类间的最佳几何角度.首先我们提出了紧致损失和等角损失促进神经坍塌的形成.其次,通过分类器向量辅助模块和难例采样模块解决不平衡数据分布导致损失无法充分施展其效果的问题.实验表明了本文方法有效诱导了神经坍塌的发生,并且超过当前最优的方法.

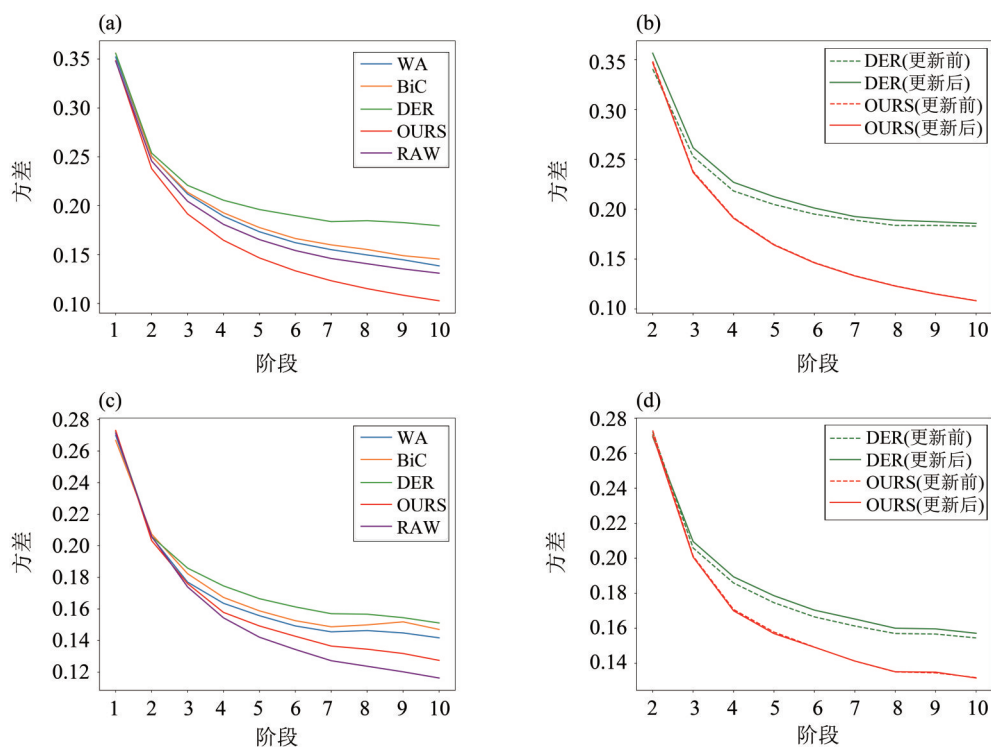


图2 类别间角度的方差

Fig. 2 Variance of angles between classes

参考文献:

- ALJUNDI R, LIN M, GOUJAUD B, et al, 2019. Gradient based sample selection for online continual learning[EB/OL]. arXiv: 1903.08671. <http://arxiv.org/abs/1903.08671>.
- BANG J, KIM H, YOO Y, et al, 2021. Rainbow memory: Continual learning with a memory of diverse samples[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 8214–8223.
- CAI T, ZHANG Z, TAN X, et al, 2023. Multi-centroid task descriptor for dynamic class incremental inference[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 7298–7307.
- CHAUDHRY A, DOKANIA P K, AJANTHAN T, et al, 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence[C]//European Conference on Computer Vision (ECCV): 556–572.
- CHEN X, CHANG X, 2023. Dynamic residual classifier for class incremental learning [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV): 18697–18706.
- CUBUK E D, ZOPH B, MANÉ D, et al, 2019. AutoAugment: Learning augmentation strategies from data[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 113–123.
- DENG J, DONG W, SOCHER R, et al, 2009. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 248–255.
- DHAR P, SINGH R V, PENG K C, et al, 2019. Learning without memorizing[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 5133–5141.
- DOUILLARD A, CORD M, OLLION C, et al, 2020. PODNet: pooled outputs distillation for small-asks incremental learning [C]//European Conference on Computer Vision (ECCV): 86–102.
- DOUILLARD A, RAMÉ A, COUAIRO G, et al, 2022. Dytox: Transformers for continual learning with dynamic token expansion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 9275–9285.
- E W, WOJTOWYTSCH S, 2020. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers[EB/OL]. arXiv: 2012.05420. <https://arxiv.org/abs/2012.05420>.
- FANG C, HE H, LONG Q, et al, 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced

- training[J]. *Proc Natl Acad Sci*, 118(43): e2103091118.
- FISHER R A, 1936. The use of multiple measurements in taxonomic problems[J]. *Ann Eugen*, 7(2): 179–188.
- FU C, DU B, ZHANG L, 2024. Do we need learnable classifiers? A hyperspectral image classification algorithm based on attention-enhanced ResBlock-in-ResBlock and ETF classifier[J]. *IEEE Trans Geosci Remote Sens*, 62: 1–13.
- GAO Q, ZHAO C, GHANEM B, et al, 2022. R-DFCIL: Relation-guided representation learning for data-free class incremental learning[C]//*European Conference on Computer Vision (ECCV)*: 423–439.
- HAN X Y, PAPYAN V, DONOHO D L, 2021. Neural collapse under MSE loss: Proximity to and dynamics on the central path [EB/OL]. arXiv: 2106.02073. <http://arxiv.org/abs/2106.02073>.
- HE K, ZHANG X, REN S, et al, 2016. Deep residual learning for image recognition[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 770–778.
- HINTON G, VINYALS O, DEAN J, 2015. Distilling the knowledge in a neural network[EB/OL]. arXiv: 1503.02531. <https://arxiv.org/abs/1503.02531>.
- HUANG B, CHEN Z, ZHOU P, et al, 2023. Resolving task confusion in dynamic expansion architectures for class incremental learning[C]//*Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 37(1): 908–916.
- HOU S, PAN X, LOY C C, et al, 2019. Learning a unified classifier incrementally via rebalancing [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 831–839.
- JI W, LU Y, ZHANG Y, et al, 2021. An unconstrained layer-peeled perspective on neural collapse[EB/OL]. arXiv: 2110.02796. <http://arxiv.org/abs/2110.02796>.
- KANG M, PARK J, HAN B, 2022. Class-incremental learning by knowledge distillation with adaptive feature consolidation[C]//*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 16050–16059.
- KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al, 2017. Overcoming catastrophic forgetting in neural networks[J]. *Proc Natl Acad Sci USA*, 114(13): 3521–3526.
- KRIZHEVSKY A, 2009. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto.
- LEE J, HONG H G, JOO D, et al, 2020. Continual learning with extended kronecker-factored approximate curvature[C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 8998–9007.
- LI Z, HOIEM D, 2018. Learning without forgetting[J]. *IEEE Trans Pattern Anal Mach Intell*, 40(12): 2935–2947.
- LI Z, SHANG X, HE R, et al, 2023. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier[C]//*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*: 5319–5329.
- LU J F, STEINERBERGER S, 2022. Neural collapse under cross-entropy loss[J]. *Appl Comput Harmon Anal*, 59: 224–241.
- LUO Z, LIU Y, SCHIELE B, et al, 2023. Class-incremental exemplar compression for class-incremental learning [C]//*2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 11371–11380.
- MIXON D G, PARSHALL H, PI J, 2022. Neural collapse with unconstrained features [J]. *Sampl Theory Signal Process Data Anal*, 20(2): 11.
- PAPYAN V, HAN X Y, DONOHO D L, 2020. Prevalence of neural collapse during the terminal phase of deep learning training [J]. *Proc Natl Acad Sci*, 117(40): 24652–24663.
- PARK W, KIM D, LU Y, et al, 2019. Relational knowledge distillation[C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 3962–3971.
- REBUFFI S A, KOLESNIKOV A, SPERL G, et al, 2017. iCaRL: Incremental classifier and representation learning[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2001–2010.
- REN J, YU C, SHENG S, et al, 2020. Balanced meta-softmax for long-tailed visual recognition[C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*: 4175–4186.
- ROMERO A, BALLAS N, KAHOU S E, et al, 2014. FitNets: Hints for thin deep nets[EB/OL]. 1412.6550. <http://arxiv.org/abs/1412.6550>.
- TAO X Y, CHANG X Y, HONG X P, et al, 2020. Topology-preserving class-incremental learning[C]//*European Conference on Computer Vision (ECCV)*: 254–270.
- WANG F Y, ZHOU D W, LIU L, et al, 2023. BEEF: Bi-compatible class-incremental learning via energy-based expansion and

- fusion[C]//The 11th International Conference on Learning Representations (ICLR): 37860–37879.
- WANG F Y, ZHOU D W, YE H J, et al, 2022. FOSTER: Feature boosting and compression for class-incremental learning[C]//European Conference on Computer Vision (ECCV): 398–414.
- WU Y, CHEN Y, WANG L, et al, 2019. Large scale incremental learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 374–382.
- YAN S, XIE J, HE X, 2021. DER: Dynamically expandable representation for class incremental learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 3014–3023.
- YANG Y, CHEN S, LI X, et al, 2022. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems: 37991–38002.
- YANG Y, ZHOU D W, ZHAN D C, et al, 2023a. Cost-effective incremental deep model: Matching model capacity with the least sampling[J]. IEEE Trans Knowl Data Eng, 35(4): 3575–3588.
- YANG Y B, YUAN H B, LI X T, et al, 2023b. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning[EB/OL]. arXiv: 2302.03004. <https://arxiv.org/abs/2302.03004>.
- ZENKE F, POOLE B, GANGULI S, 2017. Continual learning through synaptic intelligence [C]//Proceedings of the 34th International Conference on Machine Learning: 3987–3995.
- ZHAO B, XIAO X, GAN G, et al, 2020. Maintaining discrimination and fairness in class incremental learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 13205–13214.
- ZHAO H, WANG H, FU Y, et al, 2022. Memory-efficient class-incremental learning for image classification [J]. IEEE Trans Neural Netw Learn Syst, 33(10): 5966–5977.
- ZHOU J, LI X, DING T, et al, 2022. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features[C]//Proceedings of the 39th International Conference on Machine Learning: 27179–27202.

(责任编辑 冯兆永)