DER 算法复现——《机器学习》期末大作业

熊洲洲 21311392　　刘元昊 21311223

1　背景介绍

在现实生活中，人类能够轻易从历史经验和逐步学习的新概念中积累视觉知识。受此启发，类增量学习（Class-Incremental Learning）问题旨在设计一个模型能够在不遗忘旧类别的前提下，正确识别和适应随时间动态增加的新类别。这一能力对于许多应用领域的图像分类任务非常重要，例如，在人脸识别系统中，摄像头不断拍摄并识别用户的人脸图像；在医学影像诊断中，通过医疗设备产生的影像数据进行病患分类；在智能交通系统中，监控摄像头持续捕获并辨识道路上的车辆图像。然而，现代机器视觉识别系统还远远不能达到人类增量学习的水平，其面临的一个主要挑战是可塑性与稳定性困境（Stability-plasticity dilemma[1]），具体来说，过度的可塑性往往会导致灾难性遗忘（Catastrophic Forgetting）问题，而过度的稳定性又会阻碍模型适应新出现的类别。

Figure 1: 增量学习算法发展图

我们复现的论文是 Yan et al. 于 2021 年发表的《DER:Dynamically Expandable Representation for Class Incremental Learning》[2]。DER 作为图 1 中动态网络的代表性算法，主要贡献有以下三个方面：

- 开发了一种动态可扩展的表示法和一种基于类增量学习的两阶段学习方法。
- 提出了一个辅助损失步骤来促进新添加的特征模块以有效地学习新出现的类别。
- 引入了一个模型剪枝步骤来学习紧凑的特征，以减少模型的参数。

2 问题建模与算法分析

2.1 问题建模

定义 1：类增量学习问题定义 设有若干任务组 $\{D^1, D^2, \cdots, D^B\}$，则定义第 t 个时刻的任务组 $D_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$ for $y_i^t \in Y_t$，其中 y_i 对应 x_i 的标签，Y_t 是标签空间，B 是任务组的数量。设 $y_t = D_1 \cup D_2 \cup \cdots \cup D_t$，表示前 t 个任务组的并集。

类增量学习问题的目标是：在第 t 个时刻的任务组训练后，得到一个分类模型 $f(x) : X \rightarrow (Y_1 \cup \cdots \cup Y_t)$，使 $f(x)$ 在前 t 个时刻的任务组并集 y_t 上的期望正确率最高。

根据定义 1，模型在第 t 个时刻的任务组的训练阶段只能访问到对应的数据集 D^t，从而保护数据中可能涉及的用户隐私，并释放存储空间。为将模型部署至嵌入式处理器，这种不记录历史数据的限制是必要的。不过，一些方法放宽了这种限制，允许模型保存一个规模相对较小的历史数据集合，称为范例集。范例集在后续训练中可以再次提供给模型学习，以减少灾难性遗忘。

定义 2：范例集 考虑 M_t 是来自以前的任务 $M_t = \{(x_j, y_j)\}_{j=1}^{M}$ 的实例的额外集合。在范例集的帮助下，模型可以利用 $M_t \cup D_t$ 在每个任务中进行再次训练。

定义 3：类重叠假设 典型的 CIL 假设 $Y_t \cap Y_{t'} = \emptyset$ for $t \neq t'$，即不同的图像分类任务中不会出现重复的类。这一假设简化了类增量学习任务的复杂性，使得更新模型和保留知识更为简单。如果假设不同任务的类别是不重复的，那么在模型接收新任务时，可以简单地将新任务的类别添加到模型中，而不必担心与已有任务的类别重叠或冲突。

2.2 算法分析

如图 2 所示，DER 通过为每个新任务动态扩展新的网络结构来缓解类增量学习中的灾难性遗忘问题。它的核心思想是为每个新的增量任务引入一个新的模型分支，并保持旧的分支冻结，以此来保留之前任务的知识。这种方法允许网络为每个任务学习特定的特征，而不会影响到旧任务的特征表示。以下将介绍 DER 算法关键技术的细节。

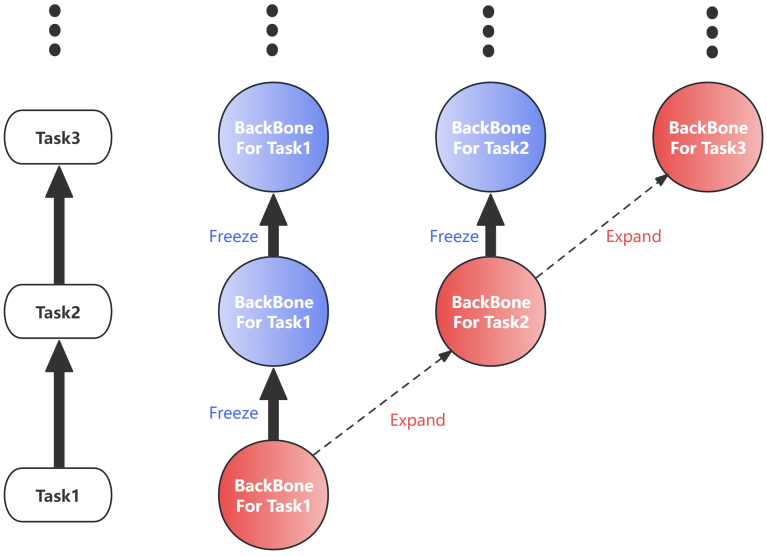


Figure 2: DER 框架

两阶段学习法

表示学习阶段 为了更好地实现稳定性和可塑性之间的权衡，DER 修正了之前的特征表示方式，并使用一个新的特征提取器对新增数据和记忆数据进行训练，从而扩展原有的特征表示。为了让这个新增的特征提取器能够更好地学习多样化且具有较强区分能力的特征，他们在这个特征提取器上设计了一个辅助损失机制。同时为了进一步提高模型的效率，他们通过引入一种基于信道级掩码的剪枝方法，实现了随新类复杂度动态扩展的表示方法。如图 3 所示。

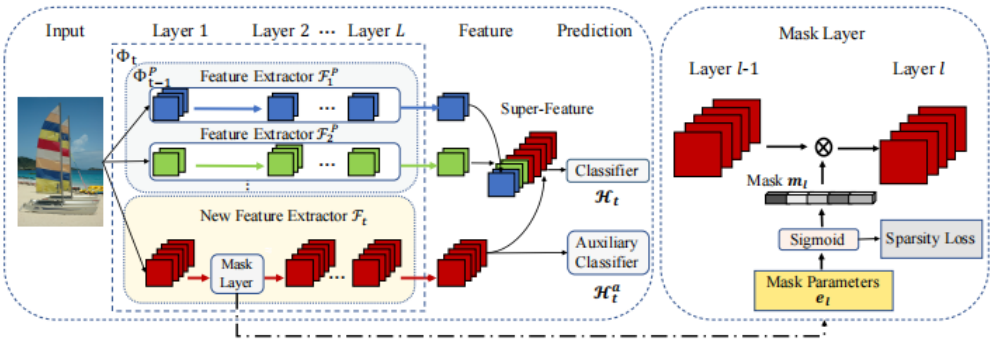


Figure 3: 动态可扩展的表示法学习

分类器学习阶段 在进行了上一阶段的表示学习后，DER 使用第 t 时刻的记忆数据 M_t 和新增数据 D_t 对分类器再次进行训练，以处理类别不平衡问题。

损失函数的设计 DER 模型最终的损失函数为 $\mathcal{L}_{DER} = \mathcal{L}_{H_t} + \lambda_a \mathcal{L}_{H_t^a} + \lambda_s \mathcal{L}_s$，其中 $\mathcal{L}_{H_t}$ 是训练损失，$\mathcal{L}_{H_t^a}$ 是辅助损失，$\mathcal{L}_s$ 是 Sparsity 损失。

训练损失 DER 模型在对记忆数据和新增数据进行训练时采用的损失函数是公式 1 的交叉熵损失 $\mathcal{L}_{H_t}(x)$。其中 $\Phi_t(x)$ 是超级特征提取器，$\mathcal{F}_t(x)$ 是新增的特征提取器，t 时刻的训练任务组 $\tilde{D}_t = M_t \cup D_t$。

$$\begin{aligned} u = \Phi_t(x) &= [\Phi_{t-1}(x), \mathcal{F}_t(x)] \\ p_{H_t}(y|x) &= Softmax(H_t(u)) \\ \mathcal{L}_{H_t} &= -\frac{1}{\left|\tilde{D}_t\right|} \sum_{i=1}^{\left|\tilde{D}_t\right|} \log p_{H_t}(y = y_i|x_i) \end{aligned} \tag{1}$$

辅助损失 为了让这个新增的特征提取器 $\mathcal{F}_t(x)$ 能够更好地学习多样化、区分强的特征，他们在训练损失的基础上额外使用了一个辅助分类器 H_t^a 设计了一个辅助损失 $\mathcal{L}_{H_i^a}$ 如公式 2 所示，并在最终的损失函数中给予一个权重 λ_a，是一个控制辅助分类器效果的超参数（在 $t = 1$ 时，$\lambda_a = 0$）。

$$\begin{aligned} p_{H_t^a}(y|x) &= Softmax(H_t^a(\mathcal{F}_t(x))) \\ \mathcal{L}_{H_t^a} &= -\frac{1}{\left|\tilde{D}_t\right|} \sum_{i=1}^{\left|\tilde{D}_t\right|} \log p_{H_t^a}(y = y_i|x_i) \end{aligned} \tag{2}$$

Sparsity 损失 他们的剪枝策略是基于可微的信道级掩码，其计算公式如 3 所示，其中 s 使用线性退火策略更新，e_l 是可学习的掩码参数，激活函数 $\sigma(\cdot)$ 采用 sigmoid 函数，信道掩码 m_l 能够控制第 l 个卷积层的大小，其范围在 $[0, 1]$ 区间。

$$\begin{aligned} s &= \frac{1}{s_{max}} + (s_{max} - \frac{1}{s_{max}})\frac{b-1}{B-1} \\ m_l &= \sigma(se_l) \end{aligned} \tag{3}$$

为了使模型在最大限度地减少参数量的同时，最小地降低模型性能。在每一个时间步中，他们添加了一个 Sparsity 损失，该损失表示已使用权重与所有可用权重的比例，如公式 4 所示，其中 L 表示卷积层的层数，K_l 表示第 l 个卷积层的卷积核大小，$\|\cdot\|_1$ 为 1 -范数（当 $l = 0$ 时对应输入图像，$\|m_0\|_1$=3），c_l 表示第 l 个卷积层的通道数，同样在最终的损失函数中赋予一个权重 λ_s 来控制模型大小。

$$\mathcal{L}_s = \frac{\sum_{l=1}^{L} K_l \left\|m_{l-1}\right\|_1 \left\|m_l\right\|_1}{\sum_{l=1}^{L} K_l c_{l-1} c_l} \tag{4}$$

3 实验评估

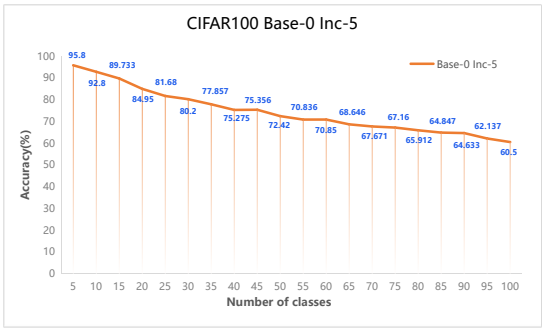
3.1 实验设置

我们小组在对 DER 算法进行实验复现时采用的数据集是 CIFAR-100。CIFAR-100 是一个 32 × 32 像素的彩色图像的数据集，共有 100 个类别，其中训练图像有 50 000 张，每个类别 500 张，测试图像有 10 000 张，每个类别 100 张。

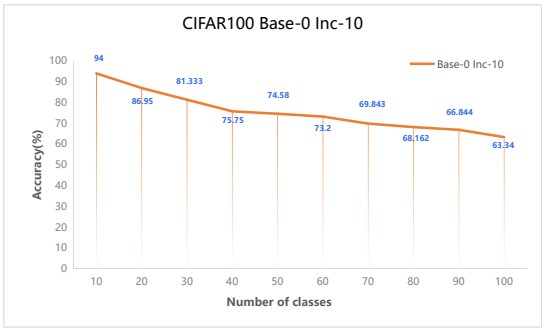
这里有两种被 CIL 研究学者广泛认可的数据集划分方式：从头开始训练（Train from scratch，TFS[3]）和从半数开始训练（Train from half，TFH[4]）。我们了规定划分准则结构为“Base-m，Inc-n”，其中 m 表示第一个训练阶段的类的数目，n 表示每个增量任务增加的类的固定数目。以下我们将分别采用“Base-0，Inc-5”、“Base-0，Inc-10”、“Base-50，Inc-10”这三种增量策略来进行实验复现，并且将我们复现的实验结果与原始论文的实验结果进行对比，同时横向对比其他算法的实验结果。

3.2 实验结果

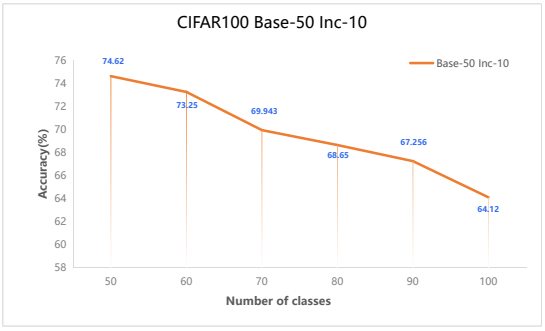
我们在论文作者发布的代码的基础上，进行了一定的参数调优（直接运行所给代码并不能得到与论文数据一致的实验结果），得到了图 4 和 5 的实验结果，并且与其他算法进行性能对比得到图 6。



(a) Base0 Inc5



(b) Base0 Inc10



(c) Base50 Inc10

Figure 4: CIFAR100 实验结果

通过图 4 我们可以观察到，这三种增量策略的实验结果的准确率整体趋势，都是随着增量任务的增加而下降，这符合类增量学习任务所面临的遗忘现象，即模型在学习新类的同时会不可避免地遗忘旧类的现象。对比图 4(a) 和 4(b) 可以发现，当每个时间步的增量类别增加时，模型能够在一个时间步内学习到更多的类别、更完整的表示，从而减少因增量次数增加导致旧类的遗忘，模型的准确率有所提高。对比图 4(b) 和 4(c) 可以得出以下结论，当第一个训练阶段的类的数目增加时，模型需要一次性学习更复杂的表示，在首个任务的准确率上会更低，但随着增量任务的进行，两种策略的准确率的差距会逐渐减少，甚至图 4(c) 的准确率会在某个时间步后反超 4(b)。

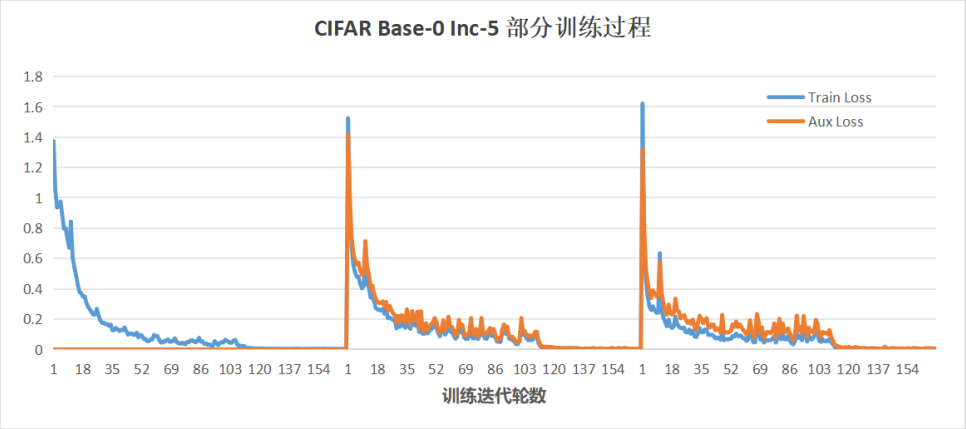


Figure 5: 训练阶段误差曲线图

图 5 描述的是增量策略 “Base-0 Inc-5” 前 3 个增量任务的训练阶段的误差，其中每个增量任务训练 170 轮，Train Loss（蓝色）是总误差中的训练损失，Aux Loss（橙色）是总误差中的辅助损失，由于复现时并没有使用剪枝策略，训练过程中不会有 Sparsity 损失。根据辅助损失的定义，第一个增量任务的 λ_a 为 0，所以在第一个增量任务的训练过程中 Aux Loss 始终为 0。在每个增量任务开始时两类损失陡增是因为有新类加入，两类损失在每个增量任务的第 110 轮训练后基本趋近于 0，表明模型能够逐渐收敛，可以学习到区分能力较强的特征。

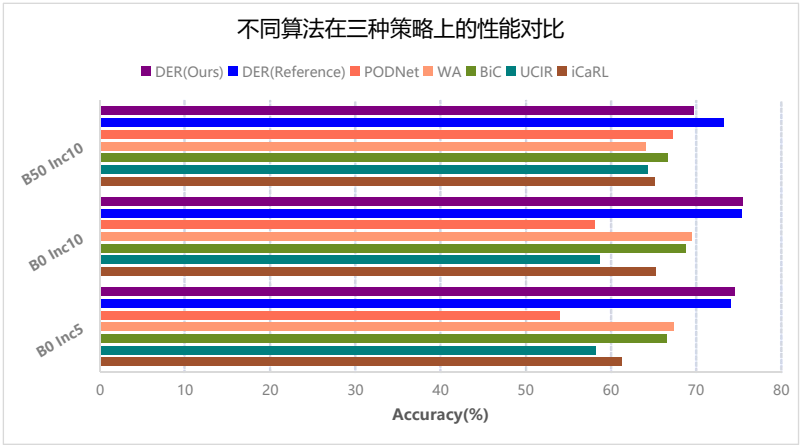


Figure 6: 算法性能对比图

在图 6 中，将我们复现的 DER 模型与作者发表数据以及作者收集的其他模型进行性能对比，具体数值可参考表 1。DER 模型在三种增量策略上始终优于所列其他模型，展现出优异的性能。我们复现的 DER 模型在“Base-0 Inc-5”和“Base-0 Inc-10”上与论文数据基本一致，甚至略微高于论文数据；在“Base-50 Inc-10”上，我们复现的结果并没有论文数据好，这可能是“Base”发生变化时，模型参数需要进一步的调整，但复现结果仍然优于所列其他模型的准确率。

Table 1: 算法性能对比表

Methods	Base-0 Inc-5	Base-0 Inc-10	Base-50 Inc-10
	Accuracy(%)		
iCaRL[3]	61.2	65.27	65.06
UCIR[4]	58.17	58.66	64.28
BiC[4]	66.48	68.8	66.62
WA[5]	67.33	69.46	64.01
PODNet[6]	53.97	58.03	67.25
DER[2]	74.09	75.36	**73.21**
Ours	**74.46**	**75.4**	69.64

论文作者还对文中新提出的扩展表示法和辅助损失进行了消融实验，得到表 2。我们可以看到，加入扩展表示法后，对模型产生了显著的积极作用，平均准确率由 61.84% 大幅提高到 73.26%，极大地提高了模型性能。在使用扩展表示法的基础上，进一步添加辅助损失后，平均准确率也得到了一定的提高，具体提高了 2.10%。总体而言，DER 新提出的组件的确能够改善模型性能，相比于原有模型平均准确率共提高了 13.52%。

Table 2: 消融实验

组件		最后阶段准确率	平均准确率
扩展表示	辅助损失		
×	×	40.81	61.84
√	×	63.07	73.26
√	√	65.34	75.36

4 结论

DER 提出了一种动态可扩展的表示法，以改进先前在类增量学习领域中的传统表示方法。在每一步的训练过程中，该模型会冻结前一个任务所学到的表示，并通过新增参数化特征增强其表示能力。此外，DER 引入了基于信道级掩码的剪枝策略，依据新类别的复杂度和辅助损失动态扩展表示法，从而更好地学习新的判别特征，同时还能减少

模型的参数数量。我们对 DER 在三种增量学习任务上进行了实验复现，进一步验证了 DER 的有效性，并展示了其在不同任务中的适用性，通过精细的参数调优，复现结果与原始论文中的实验结果基本一致，甚至在某些任务中略微优于原始论文，成功再现了该论文的实验结论。这表明 DER 能够有效地缓解类增量学习中的灾难性遗忘，证明了其在保持旧知识的同时学习新知识的良好能力，展现出其在类增量学习中的潜力和优势。

5 小组分工

熊洲洲：编写实验报告，制作讲解 PPT，讨论复现实验内容与报告结构；

刘元昊：复现论文代码、录制 PPT 讲解视频，讨论复现实验内容与报告结构。

References

[1] Stephen Grossberg. Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, 37:1–47, 2013. Twenty-fifth Anniversay Commemorative Issue.

[2] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023, June 2021.

[3] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017.

[4] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 86–102, Cham, 2020. Springer International Publishing.