软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2101−2119 [doi: 10.13328/j.cnki.jos.007022]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

视觉语言模型引导的文本知识嵌入的小样本增量学习*

姚涵涛[1], 余 璐[3], 徐常胜[1,2]

[1](多模态人工智能系统全国重点实验室 (中国科学院 自动化研究所), 北京 100190)

[2](中国科学院大学 人工智能学院, 北京 100049)

[3](天津理工大学 计算机科学与工程学院, 天津 300384)

通信作者: 姚涵涛, E-mail: hantao.yao@nlpr.ia.ac.cn

摘 要: 真实场景往往面临数据稀缺和数据动态变化的问题, 小样本增量学习的目的是利用少量数据推理数据知识并减缓模型对于旧知识的灾难性遗忘. 已有的小样本增量学习的算法 (CEC 和 FACT 等) 主要是利用视觉特征来调整特征编码器或者分类器, 实现模型对于新数据的迁移和旧数据的抗遗忘. 但是少量数据的视觉特征往往难以建模一个类别的完整特征分布, 导致上述算法的泛化能力较弱. 相比于视觉特征, 图像类别描述的文本特征具有较好的泛化性和抗遗忘性. 因此, 在视觉语言模型的基础上, 研究基于文本知识嵌入的小样本增量学习, 通过在视觉特征中嵌入具有抗遗忘能力的文本特征, 实现小样本增量学习中新旧类别数据的有效学习. 具体而言, 在基础学习阶段, 利用视觉语言模型抽取图像的预训练视觉特征和类别的文本描述, 并通过文本编码器实现预训练视觉特征到文本空间的映射. 进一步利用视觉编码器融合学习到的文本特征和预训练视觉特征抽象具有高辨别能力的视觉特征. 在增量学习阶段, 提出类别空间引导的抗遗忘学习, 利用旧数据的类别空间编码和新数据特征微调视觉编码器和文本编码器, 实现新数据知识学习的同时复习旧知识. 在 4 个数据集 (CIFAR-100, CUB-200, Car-196 和 miniImageNet) 上验证算法的有效性, 证明基于视觉语言模型文本知识嵌入可以在视觉特征的基础上进一步提升小样本增量学习的鲁棒性.

关键词: 小样本增量学习; 视觉语言模型; 文本知识嵌入; 类别空间引导的抗遗忘学习

中图法分类号: TP18

中文引用格式: 姚涵涛, 余璐, 徐常胜. 视觉语言模型引导的文本知识嵌入的小样本增量学习. 软件学报, 2024, 35(5): 2101–2119. http://www.jos.org.cn/1000-9825/7022.htm

英文引用格式: Yao HT, Yu L, Xu CS. Few-shot Incremental Learning with Textual-knowledge Embedding by Visual-language Model. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2101–2119 (in Chinese). http://www.jos.org.cn/1000-9825/7022.htm

Few-shot Incremental Learning with Textual-knowledge Embedding by Visual-language Model

YAO Han-Tao[1], YU Lu[3], XU Chang-Sheng[1,2]

[1](State Key Laboratory of Multimodal Artificial Intelligence Systems (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

[3](School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China)

Abstract: In real scenarios, the application often faces the problems of data scarcity and dynamic data changes. Few-shot incremental learning aims to use a small amount of data to infer data knowledge and reduce the model's catastrophic forgetting of old knowledge.

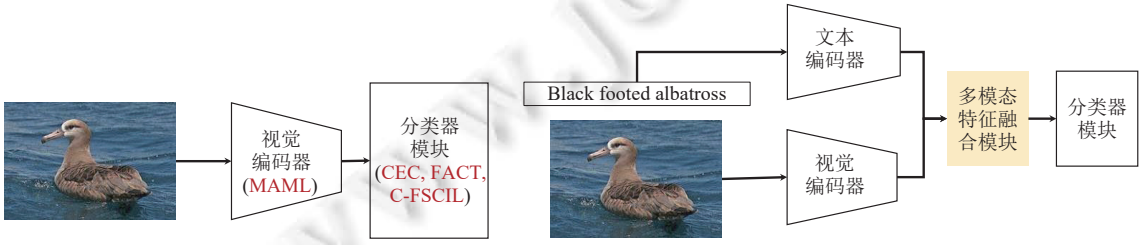
* 基金项目: 科技创新 2030—"新一代人工智能"重大项目 (2021ZD0112202); 北京市自然科学基金 (L201001, 4222039); 国家自然科学基金 (U21B2044, 62202331, 62376268)
本文由"多模态协同感知与融合技术"专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-04-06; 修改时间: 2023-06-08; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2023-11-23

Existing few-shot incremental learning algorithms (CEC, FACT, etc.) mainly use visual features to adjust the feature encoder or classifier, so as to achieve the model's transfer to new data and anti-forgetting of old data. However, the visual features of a small amount of data are often difficult to model a complete feature distribution of a class, resulting in weak generalization ability of the above algorithms. Compared with visual features, the text features of image class descriptions have better generalization and anti-forgetting abilities. Therefore, based on the visual language model (VLM), this study investigates the few-shot incremental learning based on textual knowledge embedding and realizes the effective learning of new and old class data in few-shot incremental learning by embedding text features with anti-forgetting ability in visual features. Specifically, in the basic learning stage, the study uses the VLM to extract the pre-trained visual features and class text descriptions of the image. Furthermore, the study uses the text encoder to project the pre-trained visual features to text space. Next, the study uses the visual encoder to fuse the learned text features and pre-trained visual features to abstract visual features with high discrimination ability. In the incremental learning stage, the study proposes the class space-guided anti-forgetting learning and uses the class space encoding of old data and new data features to fine-tune the visual encoder and text encoder, so as to achieve new data knowledge learning while reviewing old knowledge. This study also verifies the effectiveness of the algorithm on four datasets (CIFAR-100, CUB-200, Car-196, and miniImageNet), proving that textual knowledge embedding based on VLM can further improve the robustness of few-shot incremental learning on the basis of visual features.

Key words: few-shot incremental learning (FSIL); visual-language model; textual-knowledge embedding; class-space guided anti-forgetting learning

近年来, 随着深度学习的快速发展, 图像识别技术的效率和准确性得到了极大的进步[1,2]. 深度学习算法取得显著进步的一个主要因素是在训练时使用了大量的标注图像. 但是在现实生活中, 图像数据往往面临数据稀缺和动态变化的问题. 针对数据稀缺的问题, 小样本学习近年来已经受到了很大的关注[3–7], 它的目的是利用少量数据推理类别的知识. 在深度学习中, 数据动态变化带来的问题是模型会对旧数据产生灾难性遗忘, 例如, 模型在新类别数据上面进行训练后会遗忘之前旧类别的知识. 因此, 增量学习 (也称持续学习) 最近也得到了研究人员的重视来提升模型的抗遗忘性[8–11]. 但是小样本学习和增量学习都只能解决现实场景中的一部分问题, 因此, 它们的算法在实际部署时存在严重的泛化性弱的问题.

为了进一步提升算法对于稀缺数据和动态变化数据的鲁棒性, 最近许多研究者开始聚焦于小样本增量学习[7,12,13], 它的目的是在增量学习阶段利用少量的数据实现模型对于新类别知识的抽取, 同时还要让模型对于旧类别的数据具有抗遗忘的能力. 相比于小样本学习和传统的增量学习, 小样本增量学习具有如下两方面的挑战: 1) 新类别数据的稀疏性和新旧类别的语义鸿沟导致模型很难推理出具有辨别能力的新类别知识; 2) 新类别的稀疏性导致新旧类别知识存在严重的不平衡问题, 从而使得模型容易过拟合于新类别数据而对旧类别知识产生更严重的灾难性遗忘问题. 针对以上难点, 小样本增量学习主要通过调整视觉编码器模块或者分类器模型实现抗遗忘学习. 在视觉编码器的学习阶段, 主要通过元学习等方法实现预训练模型在当前任务的快速知识迁移[13,14]. 而在分类器模块的学习过程中主要是关联新旧类别的关系或者为新类别预留无损的特征空间[15,16]. 但是上述算法都是基于视觉特征进行小样本增量学习 (图 1(a)). 新类别数据的稀缺性导致学习到视觉特征可能与原始数据存在严重的偏差. 更进一步, 仅依赖少量数据的视觉特征是否可以构建具有抗遗忘的和高辨别性的模型也是一个值得考虑的问题.



(a) 基于视觉特征的小样本增量学习　　(b) 文本知识嵌入的小样本增量学习

图 1　基于视觉特征的小样本增量学习和基于文本知识嵌入的小样本增量学习

相比于仅有视觉特征, 视觉语言的多模态研究[17–19]表明文本特征可以提供与视觉特征描述相互补的特征. 并且, 文本语言特征具有更好的抗遗忘性因为文本单词之间存在显示的关联关系. 另外, 利用大量数据训练的视觉语言模型具有较好的泛化能力和抗遗忘能力. 基于上述考虑, 我们提出了基于文本知识嵌入的小样本增量学习 (图 1(b)). 在基类数据学习阶段, 基于类别的文本描述性, 利用视觉语言模型中的语言编码器把文本描述映射到文本特征空间. 另外, 利用视觉编码器提取图像的预训练视觉特征. 由于文本特征和预训练视觉特征都是基于预训练视觉语言模型生成的, 它们在当前任务数据集上存在一定的语义鸿沟. 因此, 我们利用文本编码器把预训练视觉特征映射到文本特征空间, 构建图像在文本空间具有辨别能力的特征描述. 进一步, 利用视觉编码器融合图像的文本特征和预训练视觉特征, 建模包含视觉和文本信息的统一特征. 在增量学习阶段, 同时考虑旧类别的空间编码和新数据的特征, 利用分类损失学习新类别和利用一致性损失复习旧知识, 实现模型对于旧类别知识的抗遗忘学习.

我们在 CIFAR-100、CUB-200、Car-196 和 miniImageNet 数据集上进行算法分析并验证了算法的有效性. 通过一系列的实验分析我们可以得出以下结论: 1) 相比于视觉特征, 文本特征在增量学习中具有更好的抗遗忘性; 2) 融合视觉信息和文本信息的统一特征可以提升小样本知识推理的准确性; 3) 旧类别空间编码辅助的增量式微调可以减少模型对于旧数据的遗忘.

1 小样本增量学习的相关工作

与小样本增量学习相关的主要工作是小样本学习、增量学习和小样本增量学习. 我们从这 3 个方面概述相关工作并总结.

(1) 小样本学习

小样本学习的目的是利用少量的标注数据推理模型实现无标注数据的识别[3,7]. 小样本学习的方法可以分为: 基于优化的算法、基于度量学习的算法、基于数据增强的方法和基于图神经网络的方法. 基于优化的算法[5,6,20–22]是利用元学习算法在少量标注数据集上实现预训练模型知识的快速迁移. 基于度量的算法[23–25]是在预训练模型的基础上通过设计度量算法来实现支持集和查询集之间的相似度学习. 在数据量有限的情况下, 基于数据增强的方法[26]通过数据增强来提高样本多样性从而提升特征的辨别能力. 由于样本数量稀缺导致难以建模样本的特征学习, 基于图神经网络的方法[27–29]利用图模型建模样本之间的关系提升数据知识的多样性达到小样本知识推理的目的.

(2) 增量学习

传统的基于静态数据学习的模型在动态场景部署时存在严重的知识灾难性遗忘问题. 在动态变化的环境中, 一个具备增量学习能力的学习系统不仅需要能够从不断出现的数据流中学习新信息, 还需要保证已学习过的信息不会发生灾难遗忘, 比如在分类任务中需要学习新的类别同时不忘记旧类别的识别能力. 增量学习算法可以分为基于记忆回放的方法、基于正则化的方法、基于动态结构的方法和基于梯度修正的方法. 基于记忆回放的方法[30–33]通常需要存储一部分旧任务的数据信息, 在学习新任务的时候, 通过回放旧任务数据与新数据结合一起训练模型, 从而达到缓解遗忘的目的. 基于正则化的方法[8–10,34]通常会设计一个额外的正则化项加入到模型目标函数中, 从而达到学习新数据更新网络参数的时候施加不同类型的约束从而达到稳定旧知识, 缓解灾难遗忘的目的. 基于动态结构的方法在学习[35–37]新任务信息的时候通过动态地改变神经网络的内部连接结构从而提供容纳新信息的空间, 从而缓解遗忘问题. 基于梯度修正的方法[38–40]在学习新任务数据的时候会对新目标函数获得的参数梯度信息进行修改, 从而保持模型对于旧任务的能力.

(3) 小样本增量学习

近年来, 小样本增量学习被提出用来解决增量场景和样本稀缺场景下的分类问题[13–16,41–46]. 现有的小样本增量学习算法可以分为: 基于元学习的算法[47–49], 基于特征空间约束的算法[15,50], 基于样本复现的算法[51]和基于动态网络结构的算法[13,16,43]. 受传统的小样本学习的启发, MetaFSCIL[47]利用元学习来迁移学习新知识并维持旧知识. 基于特征空间约束的算法目的是缓解新增样本特征空间的学习对旧特征空间的干扰. Zhou 等人[15]提出的

FACT 通过前向性兼容在基础类训练的时候为未来新类别预留分类空间从而实现旧模型在无损情形下插入新数据. 基于样本复现的算法是通过存储或者生成旧任务数据来实现新任务学习的同时复习旧任务知识, 例如, Liu 等人[51]提出了一种数据无关的伪样本生成算法并通过交叉熵正则化项来提升生成样本的多样性. 在基于动态网络结构的方法中, Zhang 等人[16]提出的 CEC 利用图模型的扩增实现不同阶段分类器知识的关联与进化, 实现新类别分类器知识有效学习的同时并保持旧类别分类器知识的有效性. Tao 等人[13]设计了神经气体网络来包括新数据和旧数据之间的拓扑关系. Ren 等人[41]利用神经气体网络来对已学习的旧知识进行抽取和表示, 并设计拓扑知识增长器来学习新知识和缓解旧知识的遗忘. 此外, 知识蒸馏也被用于小样本增量持续学习, Cheraghian 等人[52]提出了基于语义知识感知的小样本增量学习算法, 通过语义知识来关联新旧类别. Dong 等人[53]提出了基于范例关系知识蒸馏的小样本增量学习算法, 利用图的关系蒸馏有效地将旧知识迁移到新模型中.

上述小样本增量学习都是在视觉特征上增加优化约束来调整特征编码器或者分类器从而实现增量任务的动态学习. 但是在小样本增量学习中, 新增类别数据的稀缺性导致学习到新类别视觉特征与原始视觉特征存在严重的偏差, 特别是在粗粒度的类别增量学习中, 新增类别与历史类别的语义关系可能比较弱. 相比于视觉特征, 视觉语言的多模态研究表明文本特征可以提供与视觉特征描述相互补的特征. 并且, 文本语言特征具有更好的抗遗忘性由于文本单词之间存在显示的关联关系. 基于上述考虑, 在预训练的视觉语言模型基础上, 本文提出了基于文本知识嵌入的小样本增量学习, 在视觉特征中嵌入文本知识来提升特征的辨别性和抗遗忘性. 基于文本特征与视觉特征的互补性优势, 本文的算法实际上也是互补于已有的基于视觉特征的算法.

2 基础知识

(1) 小样本增量学习的形式化定义

小样本增量学习 (few-shot incremental learning, FSIL) 的目标是设计一个鲁棒的算法可以基于少量的标记数据持续学习增加的类别, 并且在不访问旧数据的情形下尽量不要对旧知识产生较大的遗忘. 标准的小样本增量学习任务往往由多个任务或者阶段组成. 假如一个小样本增量学习包含 n 个任务, 则相应的数据的定义为 $D=\{D_1,D_2,\ldots,D_n\}$, 其中 $D_i=\{x_i,y_i\}_{i=1}^{n_i}$ 代表第 i 个阶段的训练数据集, x_i 和 y_i 分别代表图像和标签, n_i 是第 i 个阶段的图像的数目. 相应地, n 个任务的测试数据定义为 $T=\{T_1,T_2,\ldots,T_n\}$. 为由于小样本增量学习往往是一个类别动态变化的增量学习任务, 因此不同阶段的类别完全不同, 即 $\forall i,j,\ Y_i\cap Y_j=\varnothing\ (i\neq j)$, Y_j 代表第 j 个任务的类别空间. 一般情况下, 为了保证增量模型可以学习到有效的知识, 第 1 个任务的数据 D_1 往往包含较多的图像数目和类别, 例如, 在 CUB-200 和 Car-196 增量学习任务中, D_1 包含了前 100 个和 96 个类别的所有训练数据. 除第 1 个任务外, 其余阶段的数据 D_t ($t>1$) 都是一个 N-way K-shot 的数据形式, 意味着每个任务包含 N 个新类别且每个类别包含 K 张图像. 小样本增量学习的目的是利用当前任务数据 D_t 推理一个对于前 t 个任务的所有类别具有高辨别能力的模型 (图 2). 它存在两方面的挑战: 一个是新增类别的图像数目较少导致难以推理有效的新类别的知识; 另一个是由于旧类别的数据在当前任务学习时是不可访问的, 如何保证模型对于旧类别知识的抗遗忘性.

(2) 视觉语言模型

视觉语言模型的目的是对齐图像特征和文本特征或者融合图像特征和文本特征, 代表性的视觉语言模型有 CLIP[54], Flamingo[55], ALIGN[56]等. 其中 CLIP 是最具有代表性的一个工作. 如图 3 所示, CLIP 模型由一个文本编码器和一个视觉编码器组成. 对于给定的图像-文本对, 利用视觉编码器提取视觉特征, 并利用文本编码器提取文本特征, 用对比损失函数来约束视觉特征和文本特征的一致性实现将两种模态数据映射到统一表达空间的目的. 由于 CLIP 是利用 4 亿个图像-文本对进行模型的训练, 训练后的模型具有很好的泛化性. 因此 CLIP 模型经常被当作预训练模型用于下游的视觉文本任务.

定义 CLIP 中的文本编码器和视觉编码器分别为 θ 和 φ. 对于一个包含 N_c 个类别的视觉文本任务. 首先利用文本编码器生成所有类别名称对应的文本空间 W. 具体而言, 给定类别名称“class”, 利用人工设计的模板“a photo of [class]”对其进行文本扩展, 并输入到文本编码器 θ 提取其对应的文本特征并构建文本语义空间 W. 对于给定的图像 x 和

其类别标签 y, 利用视觉编码器 φ 提取图像的视觉特征 $f=\varphi(x)$. 然后计算视觉特征与文本空间中每个类别的相似度:

$$p(y\mid f)=\frac{\exp(d(f,W[y])/\tau)}{\sum_{i=1}^{N_c}\exp(d(f,W[c])/\tau)} \tag{1}$$

其中, $d(\cdot)$ 是余弦相似度函数, τ 是一个可学习的温度系数.

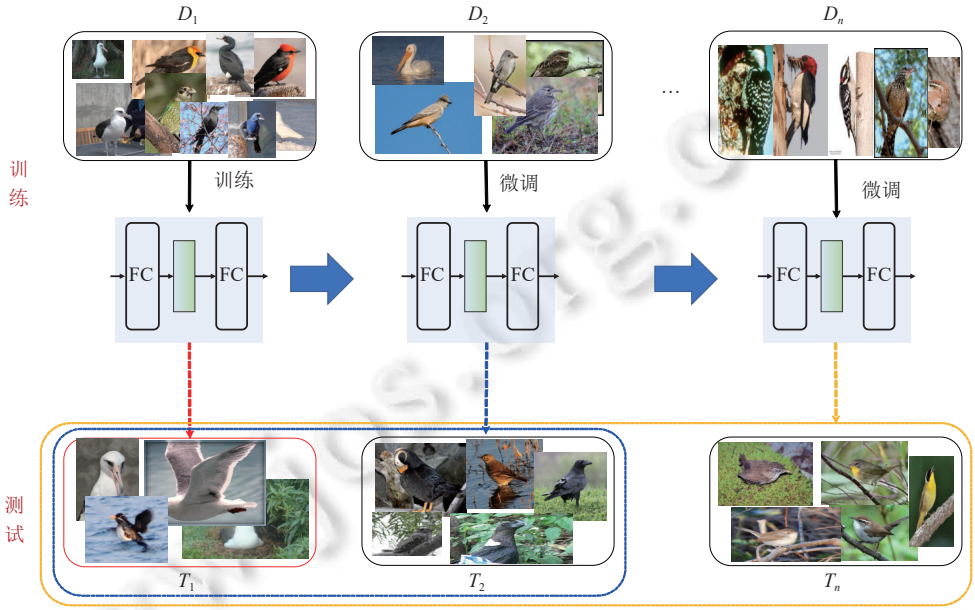


图 2 小样本增量学习的流程

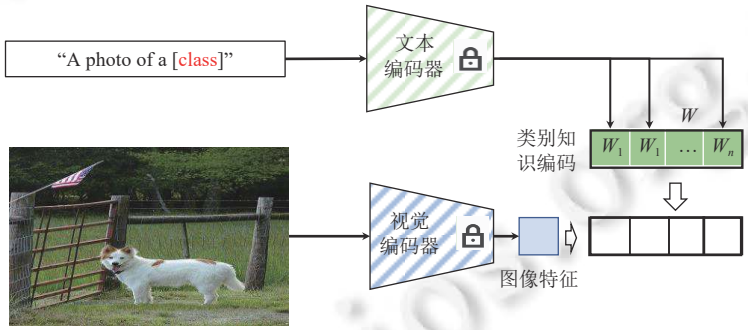


图 3 视觉语言模型 CLIP 的框架

3 文本知识嵌入的小样本增量学习

针对小样本增量学习只考虑视觉特征的问题, 我们提出了文本知识嵌入的小样本增量学习. 在样本稀缺的情形下, 学习到的视觉特征可能与图像的真实分布存在较大的偏差, 导致仅利用视觉特征难以推理具有高鲁棒性的小样本识别模型. 另外少量的视觉特征也会使得模型对于新数据的过拟合从而退化模型的泛化能力. 相比于视觉特征, 文本描述是基于单词或者词组的, 它具有天然的语义先验知识. 另外文本知识是基于人类认知的先验知识整理的, 具有较好的抗遗忘能力和泛化能力. 因此, 在小样本增量学习任务中, 通过嵌入具有先验性和抗遗忘性的文

本知识, 可以进一步提升小样本模型的推理能力. 另外, 基于大量视觉和文本数据训练的视觉语言模型已经被证明在零样本和小样本任务中具有较好的泛化能力. 因此, 在小样本增量学习中可以通过挖掘视觉语言模型中的文本知识和视觉知识来增强特征的泛化性, 从而达到缓解知识遗忘的目的.

基于上述考虑, 我们提出了基于文本知识嵌入的小样本增量学习. 如图 4 所示, 该框架包括 3 部分: 视觉语言模型、文本特征映射模块和视觉特征映射模块. 其中, 视觉语言模型用来提取具有一定泛化能力的预训练文本特征和预训练视觉特征. 需要注意的是视觉语言模型是利用预训练好的模型, 它的参数是固定的, 在增量学习的过程中不会对它进行训练和微调. 文本特征映射模型是把预训练视觉特征映射到文本空间, 而视觉特征映射模型是利用样本的文本特征和预训练视觉特征为输入来进行视觉信息和文本信息的融合, 得到嵌入了文本知识的具有高辨别能力和强抗遗忘能力的样本特征.

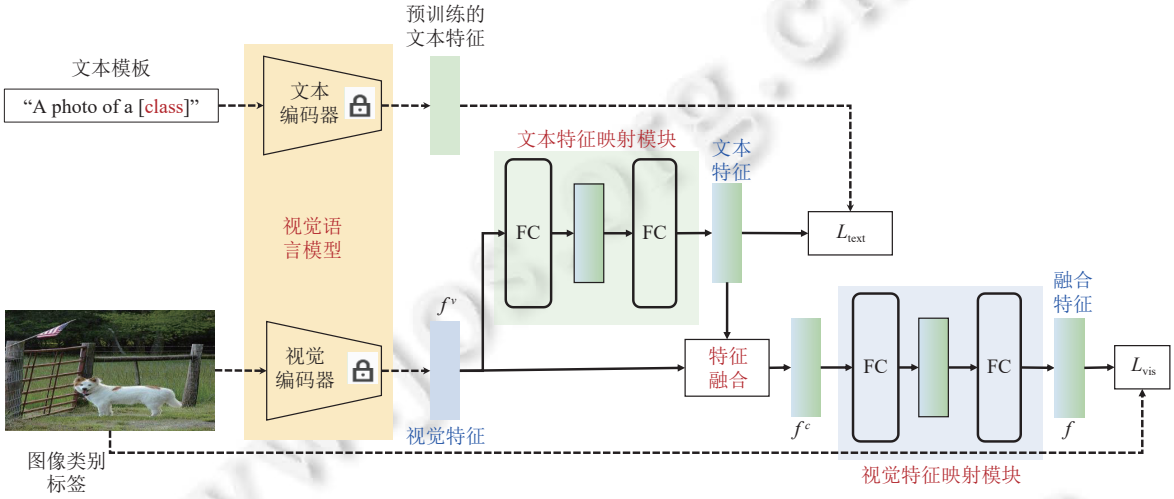


图 4 文本知识嵌入的小样本增量学习框架

由于小样本增量学习中第 1 个任务包含的图像数目和类别数目是远远大于后面的增量任务, 因此小样本增量学习可以分为两阶段: 针对第 1 个任务的文本知识嵌入的特征学习 (textual knowledge embedding representation learning, TKE) 和针对后期任务的类别空间引导的抗遗忘学习 (class-space guided anti-forgetting learning, CSG). 我们下面详细介绍这两部分的内容.

3.1 文本知识嵌入的特征学习

定义小样本增量学习中第 1 个任务的数据为 $D_1=\{x_i,y_i\}_{i=1}^{n_1}$, 其中 x_i 和 y_i 是图像和它的标签, n_1 是图像的数目. 另外, 第 1 个任务中的图像类别名称定义 $\mathbb{T}_1=\{t_c\}_{c=1}^{C_1}$, 其中 t_c 是第 c 个类别的名称, 例如 CUB-200 中第 1 个类别的名称 t_1 ={Black footed albatross}. C_1 是第 1 个任务中图像类别的数目. 给定 $\mathbb{T}_1=\{t_c\}_{c=1}^{C_1}$, 利用模板“a photo of a [class]”对每个类别名称进行扩增, 并输入到文本编码器生成对应的文本空间的特征. 因此, 对于 $\mathbb{T}_1=\{t_c\}_{c=1}^{C_1}$, 利用文本编码器可以得到对应的文本特征 $W=\{W_c\}_{c=1}^{C_1}$, $W\in\mathbb{R}^{C_1\times D}$ 可以看作是第 1 个任务类别的文本特征空间.

文本特征映射模块: 对于给定图像 x 及其标签 y, 利用视觉编码器生成对应的视觉特征 f^v. 由于视觉特征 f^v 是基于预训练模型的视觉编码器得到, 它与图像的真实特征分布和文本特征分布存在一定的偏差. 我们利用文本特征映射模块 θ_1 实现预训练视觉特征往文本空间的映射. 例如, 利用文本特征映射模块 θ_1 对视觉特征 f^v 进行映射得到对应的文本特征 $f^t=\theta_1(f^v)$. 本文中的文本特征映射模块是一个包含两个全链接层的特征映射模型 ($D_{\text{text}}\times D$), 其中第 1 层包含 D_{text} 个神经元, 实现视觉特征往高维空间的映射. 第 2 个全连接层包含 D 个神经元, 实现高维特征的降维. 在实验部分我们将分析特征维度 D_{text} 对于小样本增量学习的影响. 基于视觉语言模型得到的所有类别的文本空间 W 可以当作一个无参分类器对文本特征 f^t 计算对比损失 (公式 (2)).

$$p(y \mid f^t)=\frac{\exp(d(f^t, W[y])/\tau)}{\sum_{c=1}^{C_1}\exp(d(f^t, W[c])/\tau)} \tag{2}$$

其中, $d(\cdot)$ 是余弦相似度函数, τ 是一个可学习的温度系数. 通过优化公式 (2) 可以得到在图像在文本空间具有辨别能力的文本特征 f^t.

视觉特征映射模块: 对于图像 x, f^v 和 f^t 可以看作是预训练模型中的视觉特征和文本特征. 通过拼接特征 f^v 和 f^t 可以得到关于图像 x 融合了视觉知识和文本知识的统一特征: $f^c=[f^v, f^t]\in\mathbb{R}^{1\times 2D}$. 对于特征 f^c, 利用视觉特征映射模块 ϕ_1 进行特征的映射. 与文本特征映射模块类似, 视觉特征映射模块也包含两层 ($D/2\times D$). 与文本特征映射模块不同, 视觉特征映射模块的第 1 层是一个降维映射层. 对于特征, f^c 视觉特征映射后的特征记为 $f=\phi_1(f^c)$. 本文中定义视觉特征映射后的特征 f 为统一特征. 对于特征 f 利用分类层和交叉熵损失进行优化训练.

类别空间编码: 在第 1 个任务训练完成后, 利用训练的模型提取数据集 $D_1=\{x_i, y_i\}_{i=1}^{n_1}$ 的所有样本的视觉特征 f^v、文本特征 f^t 和统一特征 f, 记为 $\mathbb{F}_1=\{f_i^v, f_i^t, f_i\}_{i=1}^{n_1}$. 因此, 可以得到关于所有图像的 3 种空间的特征表示. CLIP 中对比学习的研究表明类别空间可以当作一个无参分类器来预测图像属于每个类别概率. 类别空间编码表示属于同一个类别的所有样本特征的平均值. 基于特征集 $\mathbb{F}_1=\{f_i^v, f_i^t, f_i\}_{i=1}^{n_1}$, 构建视觉类别空间 $M^v\in\mathbb{R}^{C_1\times D}$、文本类别空间 $M^t\in\mathbb{R}^{C_1\times D}$ 和统一类别空间 $M\in\mathbb{R}^{C_1\times D}$. 我们以统一特征 $\widehat{\mathbb{F}}_1=\{f_i, y_i\}_{i=1}^{n_1}$ 生成统一类别空间 $M\in\mathbb{R}^{C_1\times D}$ 为例介绍类别空间编码的计算过程. $M[j]$ 表示属于第 j 个类别的平均特征, 它的计算如公式 (3) 所示:

$$M[j]=\frac{1}{\mathfrak{N}(\widehat{\mathbb{F}}_1[j])}\sum\nolimits_{f\in\widehat{\mathbb{F}}_1[j]} f \tag{3}$$

其中, $\widehat{\mathbb{F}}_1[j]$ 表示特征集 $\widehat{\mathbb{F}}_1$ 中属于第 j 个类别的所有特征, $\mathfrak{N}(\widehat{\mathbb{F}}_1[j])$ 表示特征集中的特征数目. 同样可以生成文本类别空间 M^t 和视觉类别空间 M^v.

由于类别空间可以当作无参分类器, 统一类别空间 M、文本类别空间 M^t 和视觉类别空间 M^v 可以被用来预测对应特征属于每个类别的概率. 由于 3 种类别空间的特性不同, 融合这 3 个类别空间的结果来得到图像最终的类别概率. 具体而言, 给定图像 x, 利用文本知识嵌入得到其对应的特征 (f^v, f^t, f), 它的最终预测概率记为:

$$p(x)=w_1\times d(f^v, M^v)/\tau+w_2\times d(f^t, M^t)/\tau+w_3\times d(f, M)/\tau \tag{4}$$

其中, 第 1 项是视觉类别空间的预测概率, 第 2 项是文本类别空间的预测概率, 第 3 项是统一类别空间的预测概率. w_1, w_2 和 w_3 是平衡 3 种预测概率的权重. 在后面的实验中我们将验证 3 种类别空间融合的有效性和必要性.

3.2 类别空间引导的抗遗忘学习

在小样本增量学习阶段, 当前任务的数据 D_t ($t>1$) 中包含的图像数目和类别有限, 例如, 在 10-way 5-shot 情形下, 每个新增的任务仅包含 10 个类别的共 50 张图像. 由于数据量少, 仅利用这些数据微调文本特征映射模块和视觉特征映射模块会使得模型过拟合于当前任务数据, 并且对历史知识存在一定的遗忘. 因此, 小样本增量学习的增量学习阶段需要解决两个问题: 1) 基于当前任务的少量标记数据, 如何推理对于当前任务具有高辨别能力的特征? 2) 在历史任务数据不可访问的情形下, 如何在当前任务的学习过程中复习旧知识? 为了解决这两个问题, 我们提出了针对增量类别数据的类别空间引导的抗遗忘学习, 如图 5 所示.

以第 j ($j>1$) 个任务为例介绍类别空间引导的抗遗忘学习的过程. 首先定义第 $j-1$ 个任务的输出包含两个模型 (文本特征映射模块 θ_{j-1} 和视觉特征映射模块 φ_{j-1}) 和 3 个类别空间编码 (统一类别空间 M_{j-1}、文本类别空间 M_{j-1}^t 和视觉类别空间 M_{j-1}^v). $M_{j-1}\in\mathbb{R}^{O_{j-1}\times D}$ 代表第 $j-1$ 个任务输出的统一类别空间, 其中 $O_{j-1}=\sum_{z=1}^{j-1}C_z$ 代表前 $j-1$ 个任务的所有类别的数目. 给定第 j 个任务的数据 $D_j=\{x_i, y_i\}_{i=1}^{n_j}$, 第 j 个任务的目标是学习具有高辨别能力和抗遗忘能力的文本特征映射模型 θ_j 和视觉特征映射模块 φ_j:

$$\theta_j, \varphi_j=CSG(\theta_{j-1}, \varphi_{j-1}, M_{j-1}, M_{j-1}^t, M_{j-1}^v, D_j) \tag{5}$$

其中, $CSG(\cdot)$ 表示类别空间引导的抗遗忘算法.

给定当前任务的数据 $D_j=\{x_i,y_i\}_{i=1}^{n_j}$，利用 CLIP 模型中的预训练视觉编码器、文本特征映射模块 θ_{j-1} 和视觉特征映射模块 φ_{j-1} 生成当前任务数据对应的特征集 $\mathbb{F}_j=\{f_i^v,f_i^t,f_i\}_{i=1}^{n_j}$，并生成当前任务的统一类别空间 CM、文本类别空间 CM^t 和视觉类别空间 CM^v. 通过拼接旧任务的类别空间 $(M_{j-1}^v,M_{j-1}^t,M_{j-1})$ 和当前任务的类别空间 (CM^v,CM^t,CM) 可以得到新的类别空间 (M_j^v,M_j^t,M_j).

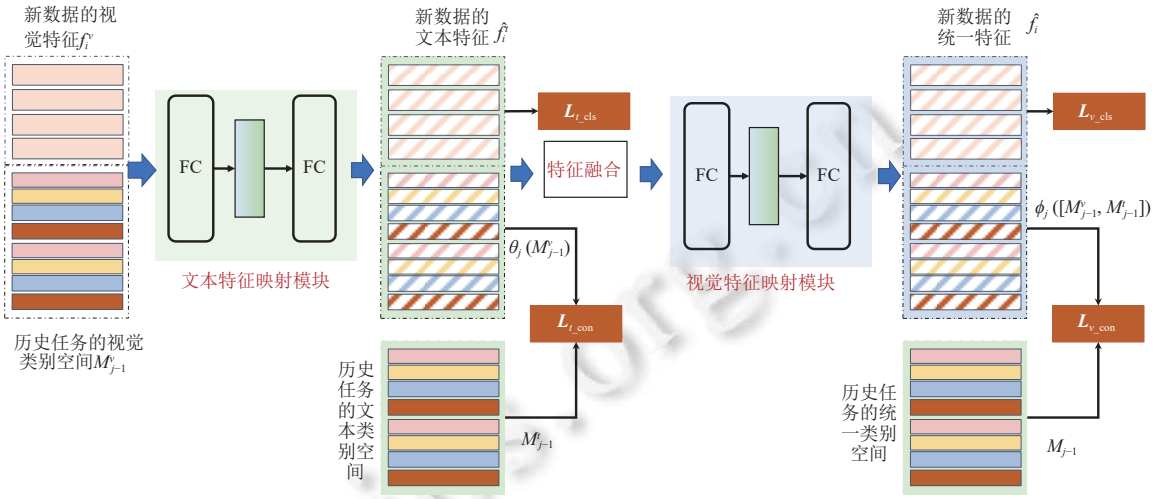


图 5 类别空间引导的抗遗忘学习

由于公式 (5) 中的优化算法 $CSG(\cdot)$ 包含两个需要优化学习的模块: 文本特征映射模块 θ_j 和视觉特征映射模块 φ_j. 我们首先讲文本特征映射模块 θ_j 的优化. 首选利用 θ_{j-1} 初始化 θ_j. 对于第 j 个任务的预训练视觉特征 f_i^v，利用文本特征映射模块 θ_j 生成其对应的文本特征 $\hat{f}_i^t=\theta_j(f_i^v)$，并利用文本类别空间计算对比损失并优化, 所有当前数据的分类损失如下:

$$L_{t\_cls}=\sum_{i=1}^{n_j}-\log\frac{\exp\left(d\left(f_i^t,M_j^t[y]\right)/\tau\right)}{\sum_{c=1}^{o_j}\exp\left(d\left(f_i^t,M_j^t[c]\right)/\tau\right)} \tag{6}$$

其中, O_j 代表前 j 个任务的类别数目总和. 利用公式 (6) 可以使得文本特征映射模块 θ_j 拟合于当前任务, 同时也会使得 θ_j 退化历史任务的知识. 为了缓解文本特征映射模块 θ_j 对于历史知识的遗忘, 进一步构造新特征和旧特征之间的一致性约束. 由于文本类别空间 M_{j-1}^t 可以看作是视觉类别空间编码 M_{j-1}^v 在旧任务的文本特征映射模块的映射. 而视觉类别空间 M_{j-1}^v 在新模型上的映射为 $\theta_j(M_{j-1}^v)$，可以构造 M_{j-1}^t 和 $\theta_j(M_{j-1}^v)$ 之间的一致性约束实现新模型对于旧知识的复习. 此外, 当前任务数据在文本特征映射模块 θ_{j-1} 上的输出也可以用于约束新旧模型在新数据上的一致性. 因此, 如公式 (7) 所示一致性约束包括两部分:

$$L_{t\_con}=\|M_{j-1}^t-\theta_j(M_{j-1}^v)\|_2^2+\sum_{i=1}^{n_j}\|f_i^t-\theta_j(f_i^v)\|_2^2 \tag{7}$$

其中, 第 1 部分是关于旧任务数据的一致性约束, 而第 2 部分是关于当前任务数据的一致性约束. 通过融合公式 (6) 和公式 (7), 文本映射模块的优化目标函数为:

$$L_t=L_{t\_cls}+w\times L_{t\_con} \tag{8}$$

其中, w 是平衡一致性约束影响的权重.

同理, 视觉特征映射模块的优化包括分类损失约束 L_{v_cls} 和一致性约束 L_{v_con}. 其中 L_{v_cls} 的定义如下:

$$L_{v\_\mathrm{cls}}=\sum_{i=1}^{n_j}-\log\frac{\exp\left(d\left(\hat{f}_i,M_j[y]\right)/\tau\right)}{\sum_{c=1}^{o_j}\exp\left(d\left(\hat{f}_i,M_j[c]\right)/\tau\right)} \tag{9}$$

其中, $\hat{f}_i=\varphi_j([f_i^v,\theta_j(f_i^v)])$ 表示视觉特征编码器对于当前特征 f_i^v 的输出. L_{v_con} 的定义如下:

$$L_{v\_\mathrm{con}}=\|M_{j-1}-\varphi_j([M_{j-1}^v,M_{j-1}^t])\|_2^2+\sum_{i=1}^{n_j}\|f_i-\hat{f}_i\|_2^2 \tag{10}$$

通过融合公式 (9) 和公式 (10), 视觉特征映射模块的优化目标如下:

$$L_v=L_{v\_\mathrm{cls}}+w\times L_{v\_\mathrm{con}} \tag{11}$$

通过融合公式 (8) 和公式 (11), 第 j 个任务下类别空间引导的抗遗忘学习的总目标函数为:

$$L=L_t+L_v \tag{12}$$

4 实验分析

4.1 实验数据

CIFAR-100[57]: CIFAR-100 是一个图像分类数据集, 它包含 60 000 张来源于 100 种类别的图像数据, 每个图像的原始大小为 32×32 . 每个类别包含 500 张训练图像和 100 张测试图像. 跟已有的小样本增量学习设置一样[13], 100 个类别中的 60 个类别的所有图像当作第 1 个任务的训练图像, 而其余的 40 个类别划分为 8 个增量任务, 每个增量任务包含 5 个类别, 每个类别包含 5 张图像.

CUB-200[58]: CUB-200 是一个关于鸟类的细粒度图像识别数据集. 它来自 200 个类别的 11 788 张图像. 每个类别平均约包含 30 张训练图像和 30 张测试图像. 与其余的小样本增量学习任务一样, 200 个类别中的 100 个类别的所有图像当作第 1 个任务的训练图像, 而其余的 100 个类别划分为 10 个增量任务, 每个增量任务包含 10 个类别, 每个类别包含 5 张图像.

Car-196[59]: Car-196 是一个关于车辆的细粒度图像识别数据集. 它共有 196 个类别, 其中的 96 个类别的所有图像当作第 1 个任务的训练图像. 而其余 100 个类别被划分为 10 个增量任务, 每个增量任务包含 10 个类别, 每个类别包含 5 张图像.

miniImageNet: miniImageNet 是一个包含 100 个类别的 ImageNet 子数据集. 其中的 60 个类别的所有图像当作第 1 个任务的训练图像. 而其余 40 个类别被划分为 8 个增量任务, 每个增量任务包含 5 个类别, 每个类别包含 5 张图像.

4.2 实现细节

我们在 CEC (https://github.com/icoz69/CEC-CVPR2021) 和 FACT (https://github.com/zhoudw-zdw/CVPR22-Fact) 提供的小样本增量学习的代码上进行修改并构建本文的算法. 在视觉语言模型中, 利用了 CLIP (ViT-L/14[60]) 作为预训练模型. 所有图像的输入都调整为 224×224, 并利用 SGD 算法进行优化训练. 在第 1 个任务的学习过程中, 训练的初始学习率是 0.005, 每隔 20 次迭代调整一次学习率, 迭代总次数是 50. 在后续增量任务的训练过程中, 训练的学习率固定为 0.001, 迭代总次数是 2 000. 图像的批次大小为 256.

由于现有的算法都是基于 ResNet[1]网络结构, 其中 ResNet18 用于 CUB-200 和 miniImageNet, ResNet20 用于 CIFAR-100. 为了与现有算法进行公平比较, 我们实现了 CLIP-ResNet18/20 网络, 其中, 利用 CLIP 预训练的文本编码器提取文本特征, 用 ResNet18/20 作为视觉编码器提取视觉特征.

4.3 评价指标及基准模型

评价指标: 在每个任务训练完成后, 利用之前任务的所有测试集进行评测分析, 并计算 Top-1 准确率. 在最后一个任务完成后, 利用性能下降率 (performance dropping rate, PD) 来评价模型的性能退化程度, $PD=A_0-A_n$, 其中 A_0 是第 1 个任务的 Top-1 准确率, A_n 是最后一个任务训练完成后的 Top-1 准确率. 此外, 我们还统计所有任务的平均准确率 $\overline{A}=\sum_{i=0}^{n}A_i$.

基准模型: 在本文的实验中, 我们复现了 CLIP[54], CEC[16]、FACT[15]和 Limit[50]等小样本增量学习的代表性方法当作一些基准模型.

● CLIP_baseline: CLIP_baseline 是直接利用视觉语言模型中的视觉编码器提取视觉特征并构建视觉类别空间编码进行预测.

● CLIP_vis: CLIP_vis 是在视觉语言模型中视觉编码器后直接利用视觉特征映射模块在第 1 个任务的数据集上进行训练, 并构建视觉类别空间编码进行预测.

● CLIP_text: CLIP_text 是在视觉语言模型中视觉编码器后直接利用文本特征映射模块在第 1 个任务的数据集上进行训练, 并构建文本类别空间编码进行预测.

● CEC: CEC 是利用元学习算法更新视觉特征映射模块, 并在分类器模块中通过图模块在旧类别的特征空间中插入新类别的特征分布.

● FACT: FACT 是在第 1 个任务的学习过程中为未来的新类别预留类别空间从而可以使得新类别分类器与旧分类器无损插入.

● Limit: Limit 通过从基础数据集中合成虚假的小样本增量任务, 并通过多任务学习和蒸馏学习提升模型的泛化能力和记忆能力.

● TKE: TKE 是本文提出的基于文本知识嵌入的小样本增量学习模型在第 1 个任务训练后的模型.

● CSG: CSG 是 TKE 在后续增量任务中利用类别空间引导的增量训练后的模型.

4.4 消融实验

● 不同类型特征的抗遗忘性分析

本文的一个研究动机是在小样本增量学习中文本知识相比于视觉特征具有更好的抗遗忘性和泛化性. 因此我们首先对比分析了预训练特征、文本特征和视觉特征等不同特征的抗遗忘性. 预训练特征是基于预训练的 CLIP 模型生成的图像特征. 视觉特征和文本特征是分别利用 CLIP_vis 和 CLIP_text 模型生成. 如图 6 所示, 相比于预训练特征和视觉特征, 文本特征具有更好的抗遗忘能力, 特别是小样本增量任务的后期任务中. 例如, 在 CUB-200 和 Car-196 数据集中预训练特征最后一个任务的 Top-1 准确率是 77.74% 和 77.9%. 在利用视觉特征映射模块对预训练特征进行训练后, 这两个数据集上最后任务的 Top-1 准确率分别提升到 78.58% 和 82.14%. 特别是对 Car-196 数据集, 相比于预训练特征, 视觉特征取得了显著性的性能提升. 利用文本特征映射模块编码后的文本特征在这两个数据集上最后任务的 Top-1 准确率是 80.43% 和 83%, 均优于预训练特征和视觉特征. 这些结果表明了文本知识在小样本增量学习中具有更好的抗遗忘性和泛化性, 同时也证明了文本知识嵌入的必要性.

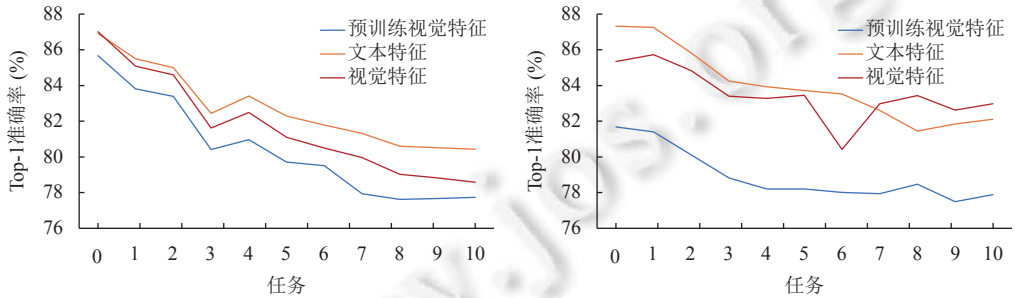


图 6 CUB-200 和 Car-196 数据集上不同类型特征的抗遗忘能力分析

● 文本特征映射模块中特征维度 D_{text} 的分析

文本特征映射模型是把预训练特征映射到文本空间, 它是一个包含两个全连接层的模型. 由于特征维度 D_{text} 会影响生成的文本特征, 因此我们在 CUB-200 数据集上分析 D_{text} 对少样本增量学习的影响. 5 种不同 D_{text} 的结果汇总在表 1 中. 可以观察到 D_{text} =1638 取得了最小的性能降低率 (PD) 和最高的平均性能 $\overline{A}$. 如表 1 所示, 后面 4

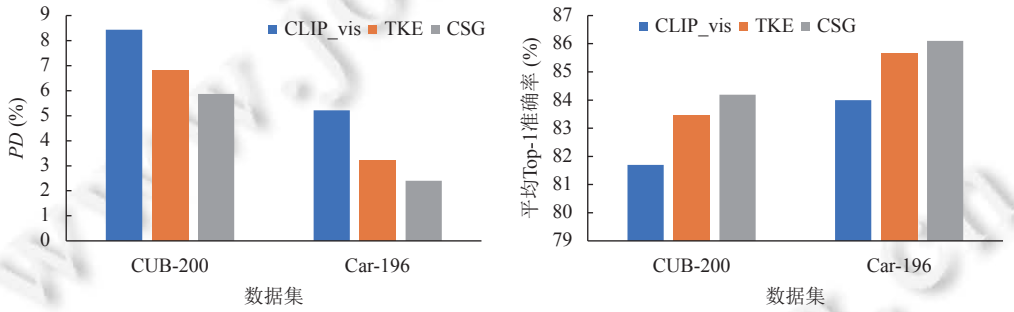
种 (768, 1 638, 1 920 和 2 304) 在第 1 个任务上的 Top-1 准确率差距不大, 表明这种设置都可以很好地在第 1 个任务上学习预训练特征到文本空间的映射. 但是这 4 种配置对后续任务的泛化性能具有不同的性能. D_{text}=768 和 D_{text}=2304 都取得了相对较差的性能, 表明它们存在较严重的遗忘能力. 因此, 本文中文本特征映射模块中 D_{text} 设为 1 638.

表 1 CUB-200 中文本特征映射模块中 D_{text} 的影响 (%)

D_{text}	Top-1											$\overline{A}\uparrow$	$PD\downarrow$
	0	1	2	3	4	5	6	7	8	9	10		
384	87.49	86.06	85.39	83.27	83.72	82.29	82.53	81.86	81.21	81.10	80.86	83.25	6.63
768	87.84	86.38	85.76	83.35	83.91	82.69	82.38	82.18	81.59	81.49	81.29	83.53	6.54
1 638	**87.94**	86.63	86.02	**84.28**	**84.82**	**83.58**	**83.17**	**83.04**	**82.32**	**82.23**	**82.07**	**84.19**	5.87
1 920	87.87	**86.75**	**86.16**	83.31	84.25	82.89	82.68	82.46	81.93	82.19	82.02	83.87	**5.85**
2 304	87.77	86.31	85.73	83.67	84.57	82.98	83.04	82.56	81.66	81.57	81.57	83.77	6.20

● 文本知识嵌入的有效性

在上述分析中已经验证了文本特征相比于视觉特征具有更好的泛化性和抗遗忘性. 本文的另一个研究动机是通过在视觉特征中嵌入文本知识达到提升辨别能力和缓解知识遗忘的目的. 因此, 我们在本节中验证文本知识嵌入对于小样本增量学习的有效性. 如图 7 所示, 在第 1 个任务的学习过程中嵌入了文本知识的 TKE 取得了优于纯视觉特征 CLIP_vis 的结果, 例如, 性能下降率 (*PD*) 从 8.43% 下降到 6.83%, 而 Top-1 的平均准确率从 81.7% 提升到 83.47%. 而通过进一步在新增的类别数据上进行类别空间引导的抗遗忘学习, CSG 可以进一步提升平均准确率和降低性能下降率.



(a) 性能下降率 (*PD*) (b) 平均Top-1准确率

图 7 文本知识嵌入的有效性分析

● CSG 模块中不同损失的分析

在类别空间引导的抗遗忘学习 (CSG) 中, 它包含文本特征映射模块的微调和视觉特征映射模块的微调, 主要是通过 L_t 和 L_v 来控制, 其中 L_t 由 L_{t_cls} 和 L_{t_con} 组成, 而 L_v 由 L_{v_cls} 和 L_{v_con} 组成. 我们因此分析一下这些损失约束的有效性. 从表 2 中可以看到, 相比于基准模型, 单一考虑 L_{t_cls}, L_{v_cls} 和 L_{v_con} 都实现了性能的提升, 例如, 利用 L_{t_cls} / L_{v_cls} / L_{v_con} 分别把平均准确率从 83.47% 提升到 83.78%/84.02%/83.59%. 其中, 分类损失约束 L_{t_cls} 和 L_{v_con} 的性能提升比较明显. 当同时优化文本特征映射模块和视觉特征映射模块的所有约束时, 算法得到了所有配置中最优的性能, 例如, 同时考虑 4 种约束得到了最低的性能下降率 5.87% 和最高的平均准确率 84.19%.

● 一致性约束权重 w 的影响

在类别空间引导的抗遗忘学习中, 公式 (8) 和公式 (11) 中权重 w 被用来平衡提升抗遗忘的一致性损失和提升辨别性的分类损失的影响. 我们因此分析不同 w 的影响. CUB-200 和 Car-196 数据集上不同权重的性能降低率 (*PD*) 和平均性能 $\overline{A}$ 的统计如图 8 所示. 其中 $w=0$ 表示不考虑一致性约束. 性能降低率越小越好, 平均性能越高越好. 我们可以看到其余的设置都取得了优于 $w=0$ 的性能, 表明了利用一致性特征约束可以提升模型的抗遗忘性. 此外, 我们也看到 $w=1000$ 在所有设置中取得了最优的性能. 因此, 在后续的实验中设置 $w=1000$.

表 2 类别空间编码引导的模型微调中不同模块的影响 (%)

L_{v_cls}	L_{v_cls}	L_{t_con}	L_{v_con}	$PD\downarrow$	$\overline{A}\uparrow$
—	—	—	—	6.82	83.47
√	—	—	—	6.33	83.78
—	√	—	—	6.00	84.02
—	—	√	—	6.82	83.47
—	—	—	√	6.65	83.59
√	√	—	—	6.45	83.61
—	—	√	√	6.56	83.64
√	—	√	—	6.42	83.72
—	√	—	√	6.10	84.10
√	√	√	√	**5.87**	**84.19**

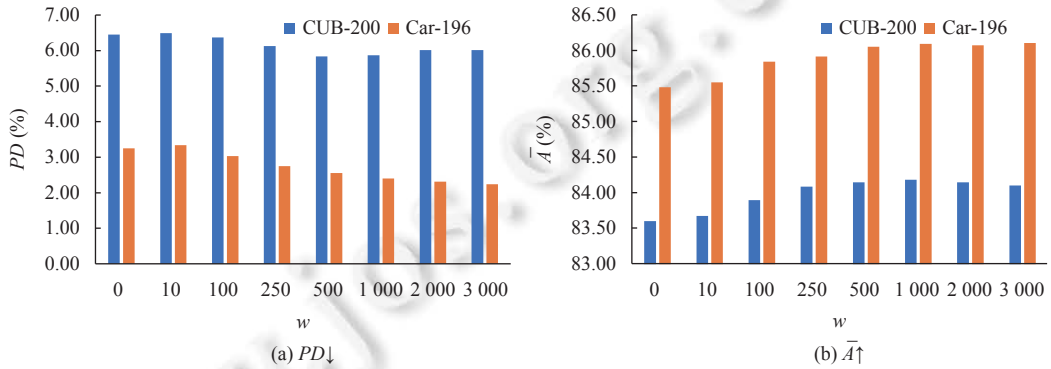


图 8 CUB-200 和 Car-196 上不同权重的性能降低率和平均性能

● 特征空间的可视化结果

为了进一步展示算法的有效性, 我们在图 9 中展示了不同类别特征在不同增量学习阶段的特征可视化结果. 首先, 我们可以观察到不同增量学习阶段的预训练模型提取的基准特征 (base feature) 的类别之间差异性较小, 类别之间融合混淆. 特别是增量学习阶段, 新增类别的特征分别很容易与基本特征的分布混淆起来, 从而降低了可分辨性. 相比于基准特征 (base feature), 我们算法中所生成的文本特征 (text feature) 和视觉特征 (visual feature) 可以显著提升特征的差异性. 从图 9 的第 3 行中可以观察到, 视觉特征的最后一个阶段 (100 base class+100 new class) 所学习到的特征具有显著的簇效应, 从而可以缓解增量学习中新特征空间对于旧任务特征空间的破坏.

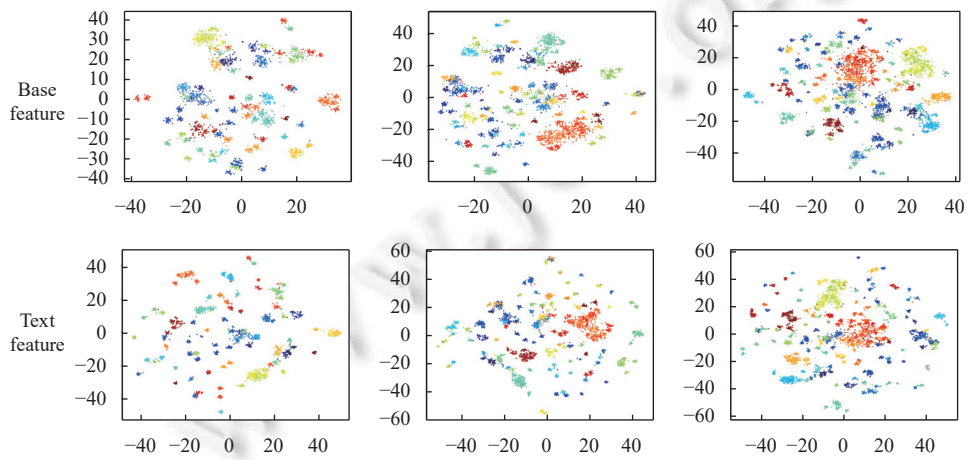
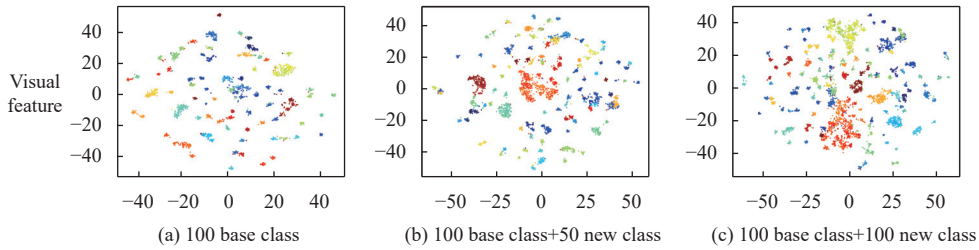


图 9 不同特征不同增量阶段的特征可视化分析



(a) 100 base class (b) 100 base class+50 new class (c) 100 base class+100 new class

图 9 不同特征不同增量阶段的特征可视化分析 (续)

● 多源类别空间融合的有效性

在公式 (4) 中, 我们通过融合 3 个类别空间的预测输出作为最终的概率输出. 因此在本节验证这种融合多元输出的有效性和合理性. 通过对 (w_1,w_2,w_3) 设置不同的值来控制不同类别空间对于最终输出的影响, 相应的结果汇总在表 3 中. 我们可以观察到融合多源类别空间的预测概率可以提升小样本增量学习的鲁棒性. 例如, 融合任意两种类别空间取得了优于单一类别空间的结果, 其中融合文本类别空间和视觉类别编码取得了较好的结果. 进一步, 融合 3 种类型的类别空间可以取得最优的结果.

表 3 不同类别空间的融合分析

w_1 (统一)	w_2 (文本)	w_3 (视觉)	CUB-200		Car-196	
			PD (%)↓	$\overline{A}$ (%)↑	PD (%)↓	$\overline{A}$ (%)↑
1.0	0.0	0.0	7.93	80.4	3.80	78.95
0.0	1.0	0.0	6.06	83.32	2.56	84.11
0.0	0.0	1.0	7.98	83.17	4.53	85.49
1.0	1.0	0.0	5.92	83.29	2.53	84.50
1.0	0.0	1.0	6.56	83.61	3.63	85.82
0.0	1.0	1.0	6.04	84.07	2.66	86.00
1.0	1.0	1.0	**5.87**	**84.19**	**2.40**	**86.10**

4.5 与现有算法的比较

在本节中, 我们在 4 个数据集 (CUB-200, CIFAR-100, Car-196 和 miniImageNet) 上将提出的算法与已有的算法进行比较分析. 基于 CLIP-ViT 的相关结果汇总在表 4–表 6 中.

表 4 CUB-200 数据集上现有算法的性能比较 (CLIP-ViT)(%)

方法	Acc in each session											PD↓	$\overline{A}$↑
	0	1	2	3	4	5	6	7	8	9	10		
CLIP_baseline[54]	85.67	83.81	83.39	80.41	80.97	79.71	79.51	77.93	77.62	77.66	77.74	7.93	80.40
CLIP_vis[54]	87.04	85.00	84.22	81.33	81.59	80.48	80.02	78.55	77.70	77.79	77.54	9.50	81.02
CEC[16]	86.87	84.96	84.16	81.30	81.76	80.71	80.23	78.80	78.12	78.16	78.01	8.86	81.19
FACT[15]	87.70	85.87	84.68	81.58	81.49	79.95	79.33	77.23	76.72	76.86	76.38	11.32	80.71
Limit[50]	87.58	85.34	84.65	81.12	81.90	80.34	79.81	78.73	78.08	78.23	77.94	9.64	81.25
TKE	87.94	86.31	85.73	83.10	84.05	83.00	82.49	82.06	81.23	81.18	81.12	6.82	83.47
CSG	**87.94**	**86.63**	**86.02**	**84.28**	**84.82**	**83.58**	**83.17**	**83.04**	**82.32**	**82.23**	**82.07**	**5.87**	**84.19**

首先, 我们可以观察到预训练的视觉语言模型 (CLIP_baseline) 在 3 个数据集上都取得了较好的基准性能, 例如, CUB-200, CIFAR-100, 和 Car-196 上 PD/$\overline{A}$ 分别是 7.93%/80.40%, 13.42%/74.73% 和 3.80%/78.95%. CLIP_baseline 的较好性能证明了视觉语言模型具有较好的泛化性和抗遗忘性, 是非常适合当作基准模型用于小

样本增量学习任务. 在 CLIP_baseline 的基础上, 加入视觉特征映射模块的 CLIP_vis 取得了更好的性能, 例如, 3 个数据集上的平均 Top-1 准确率分别从 80.40%, 74.73% 和 78.95% 提升到 81.02%, 78.37% 和 83.99%, 证明了利用视觉特征映射模块在视觉语言模型的基础上学习任务相关特征的有效性. 与这两种基准模型相比, 本文提出的文本知识嵌入的模型 (TKE) 取得了最优的性能, 例如, CUB-200, CIFAR-100, 和 Car-196 上 $PD/\overline{A}$ 结果分别是 6.82%/83.47%, 16.06%/85.66%, 和 3.22%/85.66%, 证明了通过文本知识嵌入可以学习到具有高辨别能力和强抗遗忘能力的特征. 通过对 TKE 在后续增量任务上利用类别空间引导的抗遗忘算法进行微调, CSG 进一步提升了小样本增量学习的性能.

表 5 CIFAR-100 数据集上现有算法的性能比较 (CLIP-ViT)(%)

方法	Acc in each session									$PD\downarrow$	$\overline{A}\uparrow$
	0	1	2	3	4	5	6	7	8		
CLIP_baseline[54]	82.87	79.45	77.51	74.60	73.39	72.36	72.07	70.85	69.45	13.42	74.73
CLIP_vis[54]	88.38	84.58	81.63	78.37	76.96	76.00	74.59	73.06	71.72	16.66	78.37
CEC[16]	88.38	85.12	82.70	79.57	78.46	77.36	76.50	74.81	73.33	15.05	79.58
FACT[15]	88.8	85.02	82.43	79.01	77.75	76.25	74.97	73.51	71.97	16.83	78.86
Limit[50]	88.53	85.14	82.49	79.03	77.91	77.08	76.07	76.07	73.20	15.33	79.50
TKE	89.15	85.23	82.74	79.51	78.23	77.00	75.83	74.56	73.21	16.06	79.51
CSG	**89.15**	**85.56**	**83.56**	**80.51**	**79.69**	**78.42**	**77.96**	**76.87**	**75.64**	**13.63**	**80.84**

表 6 Car-196 数据集上现有算法的性能比较 (CLIP-ViT)(%)

方法	Acc in each session											$PD\downarrow$	$\overline{A}\uparrow$
	0	1	2	3	4	5	6	7	8	9	10		
CLIP_baseline[54]	81.70	81.43	80.13	78.84	78.22	78.23	78.03	77.95	78.49	77.51	77.90	3.80	78.95
CLIP_vis[54]	87.05	86.84	85.58	84.12	83.77	83.42	83.24	82.74	83.02	82.02	82.07	4.98	83.99
CEC[16]	87.05	87.02	86.02	84.66	84.48	84.21	84.10	83.65	83.93	83.06	83.14	3.91	84.67
FACT[15]	**88.30**	**88.36**	86.83	84.91	84.44	83.90	83.56	82.95	83.05	81.98	81.92	6.38	84.56
Limit[50]	87.38	87.41	86.30	84.82	84.60	84.64	84.29	83.98	84.30	83.12	83.30	4.08	84.92
TKE	87.62	87.89	87.05	85.59	85.44	85.37	85.14	84.61	84.92	84.18	84.40	3.22	85.66
CSG	87.62	87.75	**87.18**	**85.94**	**85.75**	**85.94**	**85.59**	**85.40**	**85.75**	**85.00**	**85.20**	**2.40**	**86.10**

现有算法中, CEC, FACT 和 Limit 是 3 种具有代表性的小样本增量学习算法, 我们利用与 CSG 相同的模型结构复现这 3 种算法. 从表 4–表 6 可以发现 FACT 算法在前期的几个任务上可以取得较优的性能, 但是在后期任务中性能下降比较严重. FACT 算法取得较好性能的原因是它通过类别的混合扩增为未来的类别预留空间可以实现前期任务中具有高辨别能力特征的学习. 另外, CEC, FACT 和 Limit 都是基于视觉特征进行小样本增量学习的. 相比于这 3 种算法, 本文提出的算法在 3 个数据集上都取得了最好的 $PD/\overline{A}$, 证明了基于文本知识嵌入的小样本增量学习的鲁棒性.

由于现有的小样本增量工作都是基于 ResNet18/20 的轻量型网络结构, 为了更公平地与现有的工作比较, 我们进一步实现了基于 CLIP-ResNet (RN) 的算法并与已有工作比较, CIFAR-100、CUB-200 和 miniImageNet 的结果如表 7–表 9 所示. 一方面, 我们通过实验结果可以发现本文提出的 CSG 算法在 3 个数据集上相比于已有算法都取得了最优的平均准确率 ($\overline{A}$). 但是 CSG 的性能下降率都是要高于 Limit 算法[50], 性能下降率高的原因是我们的算法在第 1 个任务的基准数据集上的性能要远远优于 Limit[50]从而使得 CSG 有一个较差的性能下降率. 另一方面, 通过对比表 4–表 7, 我们可以发现 CLIP-RN 架构都取得了弱于 CLIP-ViT 的性能, 原因是 CLIP-ViT 是在大规模的数据上进行训练后的模型具有较好的泛化性.

上述的实验结果验证了文本知识嵌入算法在小样本增量学习任务中的有效性, 主要包括两方面: 一方面证明了在小样本增量学习中考虑文本知识嵌入的合理性; 另一方面证明了类别空间引导的抗遗忘学习的有效性.

表 7 CIFAR-100 数据集上现有算法的性能比较 (CLIP-RN20)(%)

方法	Acc in each session									$PD\downarrow$	$\overline{A}\uparrow$
	0	1	2	3	4	5	6	7	8		
iCaRL[31]	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	50.37	32.87
EEIL[61]	64.10	53.11	43.71	35.15	28.96	24.98	21.01	17.26	15.85	48.25	33.79
Decoupled-DeepEMD[25]	69.75	65.06	61.20	57.21	53.88	51.40	48.80	46.84	44.41	25.34	55.39
TOPIC[13]	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	34.73	42.62
CEC[16]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	23.93	59.53
MetaFSCIL[47]	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.47	25.03	60.73
Replay[51]	74.4	70.2	66.54	62.51	59.71	56.58	54.52	52.39	50.14	24.26	60.78
Limit[50]	73.02	70.76	67.45	**63.38**	59.97	56.90	**54.84**	52.18	49.92	**23.1**	60.94
FACT[15]	**77.13**	70.64	66.57	62.70	59.85	56.94	54.64	52.34	50.2	26.93	61.22
CSG	76.55	**72.15**	**67.53**	63.25	**60.16**	**57.05**	54.57	**52.73**	**50.2**	26.35	**61.58**

表 8 CUB-200 数据集上现有算法的性能比较 (CLIP-RN18)(%)

方法	Top-1											$PD\downarrow$	$\overline{A}\uparrow$
	0	1	2	3	4	5	6	7	8	9	10		
iCaRL[31]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	47.52	36.67
EEIL[61]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11	46.57	36.27
TOPIC[13]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.26	42.40	43.92
Decoupled-DeepEMD[25]	75.35	70.69	66.68	62.34	59.76	56.54	54.61	52.52	50.73	49.20	47.60	27.75	58.73
Replay[51]	75.90	72.14	68.64	63.76	62.58	59.11	57.82	55.89	54.92	53.58	52.39	23.51	61.52
CEC[16]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	23.57	61.33
MetaFSCIL[47]	75.90	72.41	68.78	64.78	64.96	59.99	58.30	56.85	54.78	53.82	52.64	23.26	62.11
FACT[15]	75.90	73.23	70.84	66.13	65.56	62.15	61.74	59.83	58.41	57.89	56.94	18.96	64.42
Limit[50]	75.89	73.55	71.99	**68.14**	67.42	63.61	62.40	61.35	59.91	58.66	57.41	**18.48**	65.48
CSG	**79.95**	**76.52**	**72.74**	67.84	**67.47**	**64.57**	**64.09**	**62.39**	**61.20**	**60.83**	**59.60**	20.35	**67.02**

表 9 miniImageNet 数据集上现有算法的性能比较 (CLIP-RN18)(%)

方法	Acc in each session									$PD\downarrow$	$\overline{A}\uparrow$
	0	1	2	3	4	5	6	7	8		
iCaRL[31]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	44.10	33.29
EEIL[61]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	41.73	34.97
TOPIC[13]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	36.89	39.64
Decoupled-DeepEMD[25]	69.77	64.59	60.21	56.63	53.16	50.13	47.79	45.42	43.41	26.36	54.57
CEC[16]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	24.37	57.75
Replay[51]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	23.63	58.02
Limit[50]	72.32	67.78	63.39	60.16	57.32	54.15	52.12	50.2	48.83	23.13	58.47
MetaFSCIL[47]	72.04	67.94	63.77	60.29	57.58	55.16	52.9	50.79	49.19	**22.85**	58.85
CSG	**73.30**	**68.57**	**64.13**	**60.87**	**58.27**	**55.67**	**53.00**	**50.98**	**49.47**	23.83	**59.36**

5 总 结

针对真实场景下数据稀缺和数据动态变化的问题, 小样本增量学习近来受到了极大的关注并取得了一些研究进展. 但是目前的算法都是基于图像的视觉特征进行小样本知识推理学习. 但是数据稀缺会导致推理的视觉特征与数据类别的原始分别存在严重的偏差. 相比于视觉特征, 图像类别的文本特征具有较好的抗遗忘性. 因此, 本文

提出了文本知识嵌入的小样本增量学习. 一方面, 在第一个任务的学习中通过在视觉特征中嵌入文本知识提升特征的辨别能力; 另一方面, 在后续的增量任务中, 利用类别空间引导的抗遗忘学习算法提升小样本增量任务下特征的抗遗忘性. 在 4 个数据集 (CUB-200, CIFAR-100, Car-196 和 miniImageNet) 上验证了本文提出算法的有效性. 目前的文本知识是基于图像类别名称来提取的, 缺少对于图像类别更详细的描述. 在后续的研究中, 可以通过构建每个类别的更加详细的文本描述来提升文本知识嵌入的小样本增量学习的鲁棒性.

References:

[1] He KM, Zhang XY, Ren QS, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE. 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[3] Liu Y, Lei YB, Fan JL, Wang FP, Gong YC, Tian Q. Survey on image classification technology based on small sample learning. Acta Automatica Sinica, 2021, 47(2): 297–315 (in Chinese with English abstract). [doi: 10.16383/j.aas.c190720]

[4] Du YD, Feng L, Tao P, Gong X, Wang J. Research on meta-transfer learning in cross-domain image classification with few-shot. Journal of Image and Graphics, 2023, 28(9): 2899–2912 (in Chinese with English abstract). [doi: 10.11834/jig.220664]

[5] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 1126–1135.

[6] Jamal MA, Qi GJ. Task agnostic meta-learning for few-shot learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11711–11719. [doi: 10.1109/CVPR.2019.01199]

[7] Ge YZ, Liu H, Wang Y, Xu BL, Zhou Q, Shen FR. Survey on deep learning image recognition in dilemma of small samples. Ruan Jian Xue Bao/Journal of Software, 2022, 33(1): 193–210 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6342.htm [doi: 10.13328/j.cnki.jos.006342]

[8] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D, Hadsell R. Overcoming catastrophic forgetting in neural networks. Proc. of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521–3526. [doi: 10.1073/pnas.1611835114]

[9] Lee SW, Kim JH, Jun J, Ha JW, Zhang BT. Overcoming catastrophic forgetting by incremental moment matching. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4655–4665.

[10] Aljundi R, Babiloni F, Elhoseiny M, Rohrbach M, Tuytelaars T. Memory aware synapses: Learning what (not) to forget. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 144–161. [doi: 10.1007/978-3-030-01219-9_9]

[11] Zhu F, Zhang XY, Liu CL. Class incremental learning: A review and performance evaluation. Acta Automatica Sinica, 2023, 49(3): 635–660 (in Chinese with English abstract). [doi: 10.16383/j.aas.c220588]

[12] Zhao HB, Fu YJ, Li XW, Li SY, Omar B, Li X. Few-shot class-incremental learning via feature space composition. arXiv:2006.15524, 2020.

[13] Tao XY, Hong XP, Chang XY, Dong SL, Wei X, Gong YH. Few-shot class-incremental learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12180–12189. [doi: 10.1109/CVPR42600.2020.01220]

[14] Hersche M, Karunaratne G, Cherubini G, Benini L, Sebastian A, Rahimi A. Constrained few-shot class-incremental learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE. 2022. 9047–9057. [doi: 10.1109/CVPR52688.2022.00885]

[15] Zhou DW, Wang FY, Ye HJ, Ma L, Pu SL, Zhan DC. Forward compatible few-shot class-incremental learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE. 2022. 9036–9046. [doi: 10.1109/CVPR52688.2022.00884]

[16] Zhang C, Song N, Lin GS, Zheng Y, Pan P, Xu YH. Few-shot incremental learning with continually evolved classifiers. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12450–12459. [doi: 10.1109/CVPR46437.2021.01227]

[17] Zhang HY, Wang TB, Li MZ, Zhao Z, Pu SL, Wu F. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of Image and Graphics, 2022, 27(9): 2652–2682 (in Chinese with English abstract). [doi: 10.11834/jig.220173]

[18] Yin J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. Survey on vision-language pre-training. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2000–2023 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[19] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[20] Liu YY, Schiele B, Sun QR. An ensemble of epoch-wise empirical Bayes for few-shot learning. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 404–421. [doi: 10.1007/978-3-030-58517-4_24]

[21] Park E, Oliva JB. Meta-curvature. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 298.

[22] Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[23] Gidaris S, Komodakis N. Dynamic few-shot visual learning without forgetting. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4367–4375. [doi: 10.1109/CVPR.2018.00459]

[24] Hou RB, Chang H, Ma BP, Shan SG, Chen XL. Cross attention network for few-shot classification. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 360.

[25] Zhang C, Cai YJ, Lin GS, Shen CH. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12200–12210. [doi: 10.1109/CVPR42600.2020.01222]

[26] Wang YX, Girshick R, Hebert M, Hariharan B. Low-shot learning from imaginary data. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7278–7286. [doi: 10.1109/CVPR.2018.00760]

[27] Satorras GV, Estrach BJ. Few-shot learning with graph neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[28] Kim J, Kim T, Kim S, Too CD. Edge-labeling graph neural network for few-shot learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11–20. [doi: 10.1109/CVPR.2019.00010]

[29] Gidaris S, Komodakis N. Generating classification weights with GNN denoising autoencoders for few-shot learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 21–30. [doi: 10.1109/CVPR.2019.00011]

[30] Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH. iCaRL: Incremental classifier and representation learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5533–5542. [doi: 10.1109/CVPR.2017.587]

[31] Shin H, Lee JK, Kim J, Kim J. Continual learning with deep generative replay. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 2994–3003.

[32] Wu CS, Herranz L, Liu XL, Wang YX, van de Weijer J, Raducanu B. Memory replay GANs: Learning to generate images from new categories without forgetting. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 5966–5976.

[33] Kamra N, Gupta U, Liu Y. Deep generative dual memory network for continual learning. arXiv:1710.10368, 2017.

[34] Liu XL, Masana M, Herranz L, Van de Weijer J, López AM, Bagdanov AD. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In: Proc. of the 24th Int'l Conf. on Pattern Recognition (ICPR). Beijing: IEEE, 2018. 2262–2268. [doi: 10.1109/ICPR.2018.8545895]

[35] Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, Pascanu R, Hadsell R. Progressive neural networks. arXiv:1606.04671, 2022.

[36] Yoon J, Yang E, Lee J, Hwang SJ. Lifelong learning with dynamically expandable networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[37] Rajasegaran J, Hayat M, Khan SH, Khan FS, Shao L. Random path selection for continual learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 12648–12658.

[38] Zeng GX, Chen Y, Cui B, Yu S. Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence, 2019, 1(8): 364–372. [doi: 10.1038/s42256-019-0080-x]

[39] He X, Jaeger H. Overcoming catastrophic interference using conceptor-aided backpropagation. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[40] Farajtabar M, Azizan N, Mott A, Li A. Orthogonal gradient descent for continual learning. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 3762–3773.

[41] Ren MY, Liao RJ, Fetaya E, Zemel RS. Incremental few-shot learning with attention attractor networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 5275–5285.

[42] Ayub A, Wagner AR. Cognitively-inspired model for incremental learning using a few examples. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE. 2020. 897906. [doi: 10.1109/CVPRW50498.2020.00119]

[43] Yang BY, Lin MB, Zhang YX, Liu BH, Liang XD, Ji RR, Ye QX. Dynamic support network for few-shot class incremental learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 45(3): 2945–2951. [doi: 10.1109/TPAMI.2022.3175849]

[44] Zhu K, Cao Y, Zhai W, Cheng J, Zha ZJ. Self-promoted prototype refinement for few-shot class-incremental learning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6797–6806. [doi: 10.1109/CVPR46437.2021.00673]

[45] Akyürek AF, Akyürek E, Wijaya DT, Andreas J. Subspace regularizers for few-shot class incremental learning. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[46] Tian SS, Li LS, Li WJ, Ran H, Ning X, Tiwari P. A survey on few-shot class-incremental learning. arXiv:2304.08130, 2023.

[47] Chi ZX, Gu L, Liu H, Wang Y, Yu YH, Tang J. MetaFSCIL: A meta-learning approach for few-shot class incremental learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 14166–14175. [doi: 10.1109/CVPR52688.2022.01377]

[48] Zou YX, Zhang SH, Li YH, Li RX. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: NeurIPS, 2022. 27267–27279.

[49] Yang YB, Yuan HB, Li XT, Lin ZC, Torr PHS, Tao DC. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net. 2023.

[50] Zhou DW, Ye HJ, Ma L, Xie D, Pu SL, Zhan DC. Few-shot class-incremental learning by sampling multi-phase tasks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12816–12831. [doi: 10.1109/TPAMI.2022.3200865]

[51] Liu H, Gu L, Chi ZX, Wang Y, Yu YH, Chen J, Tang J. Few-shot class-incremental learning via entropy-regularized data-free replay. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 146–162. [doi: 10.1007/978-3-031-20053-3_9]

[52] Cheraghian A, Rahman S, Fang PF, Roy SK, Petersson L, Harandi M. Semantic-aware knowledge distillation for few-shot class-incremental learning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2534–2543. [doi: 10.1109/CVPR46437.2021.00256]

[53] Dong SL, Hong XP, Tao XY, Chang XY, Wei X, Gong YH. Few-shot class-incremental learning via relation knowledge distillation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 1255–1263. [doi: 10.1609/aaai.v35i2.16213]

[54] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[55] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han TD, Gong ZT, Samangooei S, Monteiro M, Menick JL, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K. Flamingo: A visual language model for few-shot learning. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: NeurIPS, 2022. 23716–23736.

[56] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[57] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, Toronto: University of Toronto, 2009.

[58] Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-UCSD birds-200-2011 dataset. Technical Report, Pasadena: California Institute of Technology, 2011.

[59] Krause J, Stark M, Deng J, Fei-Fei L. 3D object representations for fine-grained categorization. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision Workshops. Sydney: IEEE, 2013. 554–561. [doi: 10.1109/ICCVW.2013.77]

[60] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[61] Castro FM, Marín-Jiménez MJ, Guil N, Schmid C, Alahari K. End-to-end incremental learning. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer. 2018. 241–257. [doi: 10.1007/978-3-030-01258-8_15]

附中文参考文献:

[3] 刘颖, 雷研博, 范九伦, 王富平, 公衍超, 田奇. 基于小样本学习的图像分类技术综述. 自动化学报, 2021, 47(2): 297–315. [doi: 10.16383/j.aas.c190720]

[4] 杜彦东, 冯林, 陶鹏, 龚勋, 王俊. 元迁移学习在少样本跨域图像分类中的研究. 中国图象图形学报, 2023, 28(9): 2899–2912. [doi: 10.11834/jig.220664]

[7] 葛轶洲, 刘恒, 王言, 徐百乐, 周青, 申富饶. 小样本困境下的深度学习图像识别综述. 软件学报, 2022, 33(1): 193–210. http://www.jos.org.cn/1000-9825/6342.htm [doi: 10.13328/j.cnki.jos.006342]

[11] 朱飞, 张煦尧, 刘成林. 类别增量学习研究进展和性能评价. 自动化学报, 2023, 49(3): 635–660. [doi: 10.16383/j.aas.c220588]

[17] 张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞. 视觉语言多模态预训练综述. 中国图象图形学报, 2022, 27(9): 2652–2682. [doi: 10.11834/jig.220173]

[18] 殷炯, 张哲东, 高宇涵, 杨智文, 李亮, 肖芒, 孙垚棋, 颜成钢. 视觉语言预训练综述. 软件学报, 2023, 34(5): 2000–2023. http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[19] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

姚涵涛(1989－), 男, 博士, 副研究员, CCF 高级会员, 主要研究领域为机器学习, 模式识别, 计算机视觉.

徐常胜(1969－), 男, 博士, 研究员, 博士生导师, CCF 杰出会员, 主要研究领域为多媒体分析/索引/检索, 模式识别, 计算机视觉.

余璐(1991－), 女, 副教授, 主要研究领域为机器学习, 增量学习.