软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2068−2082 [doi: 10.13328/j.cnki.jos.006771]

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

多模态引导的局部特征选择小样本学习方法*

吕天根[1], 洪日昌[1,2], 何 军[1,2], 胡社教[1]

[1](合肥工业大学 计算机与信息学院, 安徽 合肥 230031)

[2](合肥综合性国家科学中心数据空间研究院, 安徽 合肥 230036)

通信作者: 洪日昌, E-mail: hongrc.hfut@gmail.com

摘 要: 深度学习模型取得了令人瞩目的成绩, 但其训练依赖于大量的标注样本, 在标注样本匮乏的场景下模型表现不尽人意. 针对这一问题, 近年来以研究如何从少量样本快速学习的小样本学习被提了出来, 方法主要采用元学习方式对模型进行训练, 取得了不错的学习效果. 但现有方法: 1) 通常仅基于样本的视觉特征来识别新类别, 信息源较为单一; 2) 元学习的使用使得模型从大量相似的小样本任务中学习通用的、可迁移的知识, 不可避免地导致模型特征空间趋于一般化, 存在样本特征表达不充分、不准确的问题. 为解决上述问题, 将预训练技术和多模态学习技术引入小样本学习过程, 提出基于多模态引导的局部特征选择小样本学习方法. 所提方法首先在包含大量样本的已知类别上进行模型预训练, 旨在提升模型的特征表达能力; 而后在元学习阶段, 方法利用元学习对模型进行进一步优化, 旨在提升模型的迁移能力或对小样本环境的适应能力, 所提方法同时基于样本的视觉特征和文本特征进行局部特征选择来提升样本特征的表达能力, 以避免元学习过程中模型特征表达能力的大幅下降; 最后所提方法利用选择后的样本特征进行小样本学习. 在 MiniImageNet、CIFAR-FS 和 FC-100 这 3 个基准数据集上的实验表明, 所提的小样本学习方法能够取得更好的小样本学习效果.

关键词: 小样本学习; 多模态融合; 图像分类; 表示学习

中图法分类号: TP18

中文引用格式: 吕天根, 洪日昌, 何军, 胡社教. 多模态引导的局部特征选择小样本学习方法. 软件学报, 2023, 34(5): 2068–2082. http://www.jos.org.cn/1000-9825/6771.htm

英文引用格式: Lü TG, Hong RC, He J, Hu SJ. Multimodal-guided Local Feature Selection for Few-shot Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2068–2082 (in Chinese). http://www.jos.org.cn/1000-9825/6771.htm

Multimodal-guided Local Feature Selection for Few-shot Learning

LÜ Tian-Gen[1], HONG Ri-Chang[1,2], HE Jun[1,2], HU She-Jiao[1]

[1](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230031, China)

[2](Institute of Dataspace of Hefei Comprehensive National Science Center, Hefei 230036, China)

Abstract: Deep learning models have yielded impressive results in many tasks. However, the success hinges on the availability of a large number of labeled samples for model training, and deep learning models tend to perform poorly in scenarios where labeled samples are scarce. In recent years, few-shot learning (FSL) has been proposed to study how to learn quickly from a small number of samples and has achieved good performance mainly by the use of meta-learning for model training. Nevertheless, two issues exist: 1) Existing FSL methods usually manage to recognize novel classes solely with the visual features of samples, without integrating information from other modalities. 2) By following the paradigm of meta-learning, a model aims at learning generic and transferable knowledge from massive similar few-shot tasks, which inevitably leads to a generalized feature space and insufficient and inaccurate representation of sample features. To tackle the two issues, this study introduces pre-training and multimodal learning techniques into the FSL process and proposes a new multimodal-

* 基金项目: 国家自然科学基金 (61932009)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-18; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-23

guided local feature selection strategy for few-shot learning. Specifically, model pre-training is first conducted on known classes with abundant samples to greatly improve the feature representation ability of the model. Then, in the meta-learning stage, the pre-trained model is further optimized by meta-learning to improve its transferability or its adaptability to the few-shot environment. Meanwhile, the local feature selection is carried out on the basis of visual features and textual features of samples to enhance the ability to represent sample features and avoid sharp degradation of the model's representation ability. Finally, the resultant sample features are utilized for FSL. The experiments on three benchmark datasets, namely, MiniImageNet, CIFAR-FS, and FC-100, demonstrate that the proposed FSL method can achieve better results.

Key words: few-shot learning (FSL); multimodal fusion; image classification; representation learning

随着大数据时代的到来, 以深度学习为代表的人工智能技术发展迅猛, 在短视频、购物、自动驾驶、工业机器人等人们日常生活和工业生产中的应用日趋广泛. 但众所周知, 深度学习技术的成功离不开复杂的模型结构设计、大量的监督数据和高性能的计算设备, 特别是大规模的监督数据为模型学习提供了丰富的经验, 成为人工智能技术发展的基石. 然而, 实际应用中监督数据并不总是容易获得, 对训练数据的过度依赖使得典型的深度学习方法很难适应监督数据不足的学习场景. 有研究表明, 当监督数据不足以有效支撑模型训练时, 现有的深度学习模型容易呈现出严重的过拟合现象、表现出较差的模型泛化能力[1,2]. 与之相比, 人类能够轻松从少量指导信息快速学习新知识或新概念, 表现出较强的少样本学习能力, 比如一个 4 岁的孩童能很容易地将儿童画册中的"狗"和"狗熊"区分开来. 为模拟人类这种从少量监督信息快速学习新知识的能力, 近年来以研究如何从少量数据进行模型学习的小样本学习任务被提了出来, 吸引了许多研究关注.

小样本学习的难点在于: 在训练样本严重不足的情况下, 传统的模型优化方法难以被有效地用于模型优化, 这影响了人工智能技术的发展与应用. 针对这一问题, 近年来国内外大量工作在元学习 (meta-learning) 的模型优化框架下对小样本学习问题进行了研究, 包括基于度量学习的 (metric-based) 小样本学习方法[3−5]、基于模型优化的 (optimization-based) 小样本学习方法[6−9]、基于数据增广的 (Hallucination-based) 小样本学习方法[10−13]和直推式 (transductive) 小样本学习方法[14−16]. 其中, 基于度量学习的小样本学习方法试图从大量小样本任务中学习一个通用的、能够较好地捕捉样本间 (视觉) 相似关系的特征空间, 在该空间内同类样本分布较为靠近, 不同类样本之间分布较为远离. 基于模型优化的方法则通过学习一个比较好的模型初始化状态或一个较 SGD[17]等传统模型优化方法更加高效的模型优化策略, 使得模型能够在给定新类别及其少量样本的情况下通过若干次 (比如 3–5 次) 模型迭代便能成功迁移到新任务, 实现对小样本新类别的快速、高精度识别. 基于数据增广的方法主要基于 Mixup[18]、Cutout[19]、CutMix[20]、随机翻转等数据增广技术或 GAN[21]等生成模型在样本 (数据) 空间和特征空间为小样本新类别生成更多的训练样本, 将小样本学习问题转变为多样本学习问题, 从而降低模型学习的难度. 直推式小样本学习方法假设同一小样本任务中训练样本和测试样本来自同一类别空间, 方法利用测试样本中的无标签数据和训练样本中的标签数据约束模型的训练过程, 通常能够取得较好的学习效果. 尽管如此, 元学习的使用使得主流方法在小样本学习过程中过多地关注了模型对小样本新任务的快速迁移和泛化能力, 而对模型的特征表示能力关注不够. 而相对应地, 有研究表明提高模型的特征表达能力对小样本学习而言同样重要[22−24], 即获取更好的样本特征表示有助于在不影响模型泛化能力的同时进一步提高模型的小样本学习效果. 比如, Raghu 等人[22]通过对 MAML 方法[6]进行分析发现基于模型优化的小样本学习方法的学习效果主要取决于模型所学习的样本特征的质量. 受此启发, 本文从提高样本特征质量的角度出发, 重新思考了小样本学习中样本特征表示的问题, 为证明样本特征质量在小样本学习任务中的重要性提供了新的证据.

具体地, 本文将模型预训练技术和多模态学习技术引入小样本学习元学习过程, 提出一种基于多模态信息引导的局部特征选择小样本学习方法. 方法主要分为模型预训练阶段和小样本元学习两个阶段. 在第 1 阶段, 方法在大规模图像分类数据集上按照常规深度学习优化策略对模型进行预训练, 该过程赋予模型一定的初始识别能力, 能够为后续元学习提供较好的样本特征表示. 在元学习阶段, 方法基于样本之间的相似关系对样本局部特征进行选择, 这种局部特征选择操作重点突出两张图片之间相关性比较高的区域而抑制相关性较低的区域, 使得到的样本特征更具区分性, 有利于测试样本的分类. 特别地, 在上述局部特征选择的过程中方法创新性地引入了文本模态信息, 即语义词向量信息 (如: Word2Vec 或 GloVe). 文本模态信息是样本语义内容在另一种维度的表达, 相关应用在零样本学习中较为常见. 本文将其作为样本视觉特征的辅助信息应用于样本局部特征选择过程. 给定一张输入

测试图片, 区别于以往直接利用样本视觉特征进行小样本分类, 本文方法首先预测样本视觉局部特征所对应的文本特征表示, 然后基于样本的视觉特征和文本特征共同进行局部特征选择与匹配, 并基于得到的样本特征进行小样本学习. 方法在两种小样本任务设置和 3 个小样本学习标准数据集上均取得了小样本识别准确率的提升.

本文的主要贡献如下: (1) 从样本特征表示角度重新思考了小样本学习问题, 证明较好的样本特征表示能有效缓解元学习过程中由于过度关注模型泛化能力、忽略模型特征表达能力所带来的模型小样本识别精度不高的问题. (2) 从提高模型特征表达能力角度出发, 结合模型预训练技术和多模态学习技术提出一种基于多模态引导的局部特征选择小样本学习新方法. (3) 在 MiniImageNet、CIFAR-FS 和 FC-100 数据集上实验证明方法的有效性, 并系统地研究了所采用的模型预训练技术和多模态学习技术对模型小样本学习性能的影响.

1 相关工作

本文方法和基于注意力机制、多模态学习和模型预训练的现有研究之间存在相关性, 下面首先分别对这些相关内容进行简单回顾.

1.1 注意力模型

注意力机制通过模型结构或模型训练策略的设计来让模型学习到对空间、特征通道或数据之间关联性的建模能力, 其最初被用于机器翻译. 2017 年, Vaswani 等人提出 Transformer[1]使得注意力机制得到前所未有的关注, 注意力机制也因此伴随着 Transformer 的流行成为现代深度学习网络结构的重要组成部分, 如 SENet[25]、CBAM[26]和 Swin-Transformer[27]. 图神经网络[28]从结构上看也可被认为是注意力机制的一种典型应用. 小样本学习中, 对注意力机制的应用比较早的工作是 MatchingNet[3], 其中 Vinyals 等人[3]基于测试样本和训练样本之间的相关性设计了一个基于注意力机制的分类器, 在单样本设置下该分类器能够将测试样本划归为与其最相关的训练样本所在的类别. 此外, Wu 等人[29]提出双相关注意力模型来克服关系网络对物体的空间分布和纹理细节的敏感性问题. Doersch 等人[30]将 Transformer 用于小样本学习来发现测试样本和训练样本局部特征之间的空间关系, 调整后的样本特征表现出较强的跨域学习能力.

1.2 多模态学习

多模态学习 (multimodal machine learning, MMML) 是多媒体领域重要的研究内容之一, 旨在通过机器学习方法实现处理和理解多源模态信息 (如视频、图像、文本等) 的能力. 一般认为, 多模态学习可以通过利用多模态数据之间的信息互补性, 从多源数据中学到更好的特征表示, 有利于下游任务的学习. 例如, Tsai 等人[31]提出多模态 Transformer (MulT) 从非对齐的语音、视觉和文本信息中进行表示学习, 在语言分析任务上获得性能提升; Hong 等人[32]将多模态学习应用于遥感图像分类, 方法对超光谱图像和合成孔径雷达图像进行了不同的融合尝试, 做到像素级别的分类; Wang 等人[33]针对多模态学习问题提出一种全新的模型学习框架, 即通道交换网络 (channel exchanging network, CEN), 该网络能自动交换来自不同模态子网络的通道信息实现多模态融合等. 零样本学习也是一个典型的多模态学习场景. 在零样本学中, 模型通常需要结合测试样本的视觉信息和额外的属性文本描述来确定其类别信息[34,35]. 零样本学习场景是小样本学习的极端形式, 因此自然而然地可以认为利用多模态学习进行小样本学习是一个可行的研究方向[4]. 但目前多模态信息在小样本学习中的应用研究尚不充分. 2017 年, Hubert 等人[36]通过从配对的图像和文本属性以及非配对的图片和文本属性中进行多模态表示学习, 方法成功推广到零样本学习和小样本学习场景. 2019 年, Xing 等人[37]认为通过无监督形式学习得到的文本信息可以作为一种知识先验, 辅助模型从少量样本图片学习. 2020 年, Zhu 等人[38]利用样本的文本属性进行小样本图像识别, 并提出了一种属性引导的双层学习框架. 2021 年, Pahde 等人[4]设计了一种跨模态特征生成框架, 该框架能够利用图片的文本属性在特征空间生成新的视觉特征, 从而构建更加可靠的类别中心, 增强了模型的小样本分类能力.

1.3 模型预训练

基于模型预训练的迁移学习技术在图像分类[39,40]、目标检测[41]等计算机视觉任务上的应用十分广泛, 被证明可以在提升模型性能的同时有效降低模型对训练数据的需求. 比如, 早期 Sulc 等人[39]使用 ImageNet 上预训练的

GoogLeNet 模型去识别蘑菇; Ren 等人[42]在所提出的 Faster-RCNN 模型中利用 ImageNet 预训练的模型进行目标检测等. 在自然语言处理领域, BERT 模型[43]的出现也进一步证明模型预训练在表示学习上的重要作用. Zhu 等人[44]基于预训练的 BERT 模型在有监督、半监督以及无监督等多种机器翻译任务设置下都取得了可观的性能提升. Garg 等人[45]将 BERT 模型用于对抗样本生成, 方法生成的样本满足语法规则, 能够让文本分类模型性能骤降. 不仅如此, BERT 模型提出的自监督学习模式为模型训练提供了一种新的思路, 它表明在自监督学习场景下模型可以从大量语料学习一个通用的特征表示, 而该特征表示支持在下游相关任务上的快速迁移. 在 BERT 模型训练方式启发下, Chen 等人[46]和 He 等人[47]所提出的对比学习预训练过程类似地可以从大量无标签图片学习通用的特征表示, 在图像分类、目标检测、语义分割等多个下游任务上取得了媲美依赖大量标签图片通过监督学习所获得的有监督学习效果. 小样本学习早期主要基于元学习对随机初始化的模型进行训练, 近年来模型预训练技术在小样本学习中也获得了足够多的重视. 例如, 2019 年 Sun 等人[48]基于预训练的主干网络, 结合所提出的困难任务发现算法, 取得模型小样本学习能力的提升. Chen 等人[49]表明将预训练模型和余弦分类器结合得到的简单小样本学习器也能取得媲美主流元学习方法的性能. 2020 年, Chen 等人[50]提出 Meta-Baseline 模型, 发现预训练网络输出的样本特征具有很好的小样本表现, 同时还发现模型预训练和元学习之间存在互斥现象. Tian 等人[23]将预训练的模型用于知识蒸馏 (knowledge distillation), 进一步提高了模型在多个数据集上的表现, 特别地作者将利用无监督学习技术预训练得到的模型用于小样本学习任务发现该预训练的模型同样能取得不错的小样本学习效果. He 等人[24]提出了小样本学习方法 DCAP, 其中作者通过消融实验同样证实了模型预训练对模型小样本学习能力的重要影响.

本文结合多模态学习技术和模型预训练技术, 提出多模态引导的局部特征选择算法进行小样本学习, 方法在原理上与 Hou 等人[51]和 Huang 等人[52]提出的方法较为接近. 其中, Hou 等人基于样本局部特征之间的相关性进行特征选择, 但他们在局部特征选择过程仅依赖于样本的视觉特征; Huang 等人利用样本的文本属性提出一种基于注意力机制的特征选择方法, 但方法依赖于细粒度的样本文本属性描述. Pahde 等人[4]和 Zhu 等人[38]的方法与之类似, 此类方法的应用场景仅限于文本属性可获得的小样本细分类场景, 难以覆盖到更为常见的小样本学习环境. 与上述方法不同, 本文提出的局部特征选择算法能够直接利用样本的类别标签来筛选比较有代表性的局部特征, 使得同类样本特征之间具有更高的类内相关性、不同类样本特征之间具有更低的类间相关性. 此外, 由于所依赖的文本内容粒度更粗也更容易获得, 方法能够轻松胜任几种常见的小样本学习设置和数据集.

2 背景知识

2.1 *N* 类 *K* 样本学习

2003 年, Li 等人[34]在图像分类任务中首次探讨了小样本学习问题. 假设一个新类别只有一个或几个训练样本, Li 等人发现模型能够利用从其他拥有大量样本的类别学到的知识辅助小样本新类别的学习. 这种迁移学习的思想对小样本学习的影响持续至今. 但早期小样本学习主要依赖于 SIFT、HOG 等反映纹理和颜色信息的手工特征研究二分类问题, 旨在区分来自目标类别的正样本和来自非目标类别的负样本. 进入深度学习时代以后, 这种简单的二分类研究已经不能满足实际应用需求, 于是以多分类为目标任务的 N 类 K 样本 (N-way K-shot) 小样本学习被提了出来并发展成为小样本学习的主要研究对象. N 类 K 样本学习问题假设目标任务中存在 N 个新类别, 每个新类别仅包含 K 个标注样本 (K 通常取很小的值), 模型从这 $N\times K$ 个标注样本中学会识别该 N 个目标新类别. 具体地, 一个 N 类 K 样本小样本学习任务 $T=\{\mathbf{D}_{\text{train}},\mathbf{D}_{\text{test}}\}$ 由支持集 (support set) $\mathbf{D}_{\text{train}}=\{(\mathbf{x}_i^S,y_i^S)\}_{i=1}^{N_s}$ 和查询集 (query set) $\mathbf{D}_{\text{test}}=\{(\mathbf{x}_i^q)\}_{i=1}^{N_q}$ 组成, 其中支持集包含 $N_s=N\times K$ 个标签样本用于模型训练, 查询集包含 N_q 个无标签样本用于模型测试, 支持集和查询集处于相同的类别空间且有 $\mathbf{D}_{\text{train}}\cap\mathbf{D}_{\text{test}}=\varnothing$, 用 $\mathbf{C}_{\text{novel}}$ 来表示小样本任务中的 N 个新类别.

2.2 元学习

N 类 K 样本学习对于数据驱动的深度学习而言是一个非常有挑战性的学习场景. 由于训练数据的稀缺, 不论是直接模型优化还是简单的“预训练+微调”的迁移学习都难以在该场景下取得很好的效果, 例如 Kock 等人[53]和

Hoffman 等人[54]. 因此, 一段时间内小样本学习发展处于停滞状态. 2016 年, Santoro 等人[55]和 Vinyals 等人[3]在两个不同的工作中将元学习用于解决小样本学习并取得突破性进展, 小样本学习得以再次引起广泛关注. 随后, 涌现出许多基于元学习的经典小样本学习方法, 例如: Finn 等人[6]基于元学习提出 MAML 方法, 旨在学习一个支持快速迁移的模型初始化; Li 等人[7]基于元学习提出 Meta-SGD 方法, 不仅学习模型初始化还进一步学习模型更新的方向和步长; Snell 等人[5]提出 Prototypical Networks 方法基于样本和类别中心的欧式距离进行分类, 旨在学习一个泛化性能较好的特征空间; Sung 等人[56]在学习特征空间的同时学习一个通用的度量函数, 以代替欧式距离、余弦相似度量等手工设计的度量函数进行测试样本分类等. 元学习一举成为小样本环境下模型学习的主流范式.

元学习的核心思想是学会学习 (learning to learn), 其概念最早于 1987 年由 Schmidhuber 等人[57]和 Hinton 等人[58]分别在两个独立的工作中几乎同时提出. 给定目标小样本任务 $T=\{\mathbf{D}_{\text{train}},\mathbf{D}_{\text{test}}\}$, 元学习摒弃了传统的仅基于当前任务训练数据 ($\mathbf{D}_{\text{train}}$) 对模型进行定向优化的模型学习策略, 转而让模型从大量相似的小样本任务中学习如何应对该类小样本学习问题而不是某个特定任务本身. 如图 1 所示, 元学习假设在目标任务 T 之外存在一个已知数据集 $\mathbf{D}_{\text{base}}$ 包含来自 $|\mathbf{C}_{\text{base}}|$ 个类别的大量样本, 其中 $\mathbf{C}_{\text{base}}\cap\mathbf{C}_{\text{novel}}=\varnothing$, 然后从中采样大量小样本任务 $\mathbf{D}_{\text{train}}^{\text{meta}}=\{(\mathbf{D}_{\text{train}}^{i},\mathbf{D}_{\text{test}}^{i})\}_{i=1}^{N_T}$ 并利用这些小样本任务进行模型训练. 对于任意小样本任务 $\mathcal{T}_i\in\mathbf{D}_{\text{train}}^{\text{meta}}$, 它都是一个和目标小样本任务 T 相似的小样本任务. 元学习通过降低模型在这些小样本任务上的期望误差, 如公式 (1) 所示, 来提高模型应对该类小样本学习问题的能力.

$$\arg\min_{\boldsymbol{\Psi}} E_{\mathcal{T}_i\in\mathbf{D}_{\text{train}}^{\text{meta}}}[\mathcal{L}(\mathbf{D}_{\text{test}}^{i}\mid\mathbf{D}_{\text{train}}^{i},\boldsymbol{\Psi})]+\mathcal{R}(\boldsymbol{\Psi}) \tag{1}$$

其中, $\mathcal{L}(\mathbf{D}_{\text{test}}^{i}\mid\mathbf{D}_{\text{train}}^{i},\boldsymbol{\Psi})$ 为模型在小样本任务 $\mathcal{T}_i$ 上的测试误差 (通常为交叉熵损失), $\boldsymbol{\Psi}$ 为模型参数, $\mathcal{R}(\boldsymbol{\Psi})$ 为正则化项 (如权重衰减、特征正则化等). 元学习的模型优化效果可以通过模型在训练集 $\mathbf{D}_{\text{train}}^{\text{meta}}$ 上的训练误差予以反映, 但其泛化性能一般还需要在测试集 $\mathbf{D}_{\text{test}}^{\text{meta}}$ 上进行检验, 其中 $\mathbf{D}_{\text{train}}^{\text{meta}}\cap\mathbf{D}_{\text{test}}^{\text{meta}}=\varnothing$. 如果模型能在所有小样本任务 $\mathcal{T}\in\mathbf{D}_{\text{train}}^{\text{meta}}\cup\mathbf{D}_{\text{test}}^{\text{meta}}$ 都取得较好的识别效果, 我们有理由相信元学习训练后的模型在目标小样本任务 T 也能有着不错的表现.

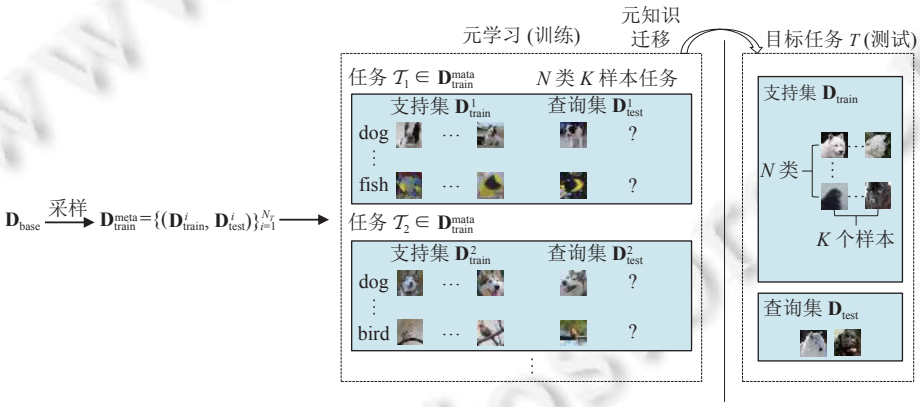

图 1 元学习任务设置和模型训练机制

3 多模态引导的局部特征选择小样本学习方法

给定已知类别数据集 $\mathbf{D}_{\text{base}}$ 和目标小样本任务 $T=\{\mathbf{D}_{\text{train}},\mathbf{D}_{\text{test}}\}$, 本文方法的流程大致如图 2 所示. 首先, 方法在已知类别数据上对模型进行常规的大规模图像分类预训练, 学习一个比较好的模型初始化状态; 相比于随机初始化, 该预训练过程显著提高模型的特征表达能力, 使得后续的元学习过程具有更快的收敛速度, 也更容易获得更好的小样本识别精度. 随后, 为了让预训练的模型能够快速适应新类别学习, 方法在元学习优化框架下对模型进行元学习; 元学习从已知类别中采样大量和目标任务 T 相似的小样本学习任务, 模型从这些小样本训练任务中学会如何应对小样本新类别学习环境. 此外, 在上述元学习阶段, 常规的做法是直接对模型输出的特征图应用全局池化

(global average pooling, *GAP*) 得到样本的特征, 并基于该特征进行小样本分类. 本文为了获得更好的小样本学习效果, 在元学习阶段使用了多模态引导的局部特征选择算法来构建具有区分性的样本特征表示. 具体地, 多模态融合模块融合样本的视觉特征和文本特征得到样本的多模态特征; 交叉注意力模块基于多模态特征计算两个样本局部特征之间的相关性; 基于得到的相关性评估, 方法对图像中相关的区域进行增强、不相关区域进行抑制, 并最终用生成样本特征计算样本之间的相似关系, 完成小样本分类. 全局视觉分类作为一种正则化手段强调在小样本训练过程中模型应保持对已知类别的识别能力, 这有效避免了模型在元学习过程中由于过于重视迁移到新任务的能力 (即泛化能力) 所造成的模型特征表达能力的衰退.

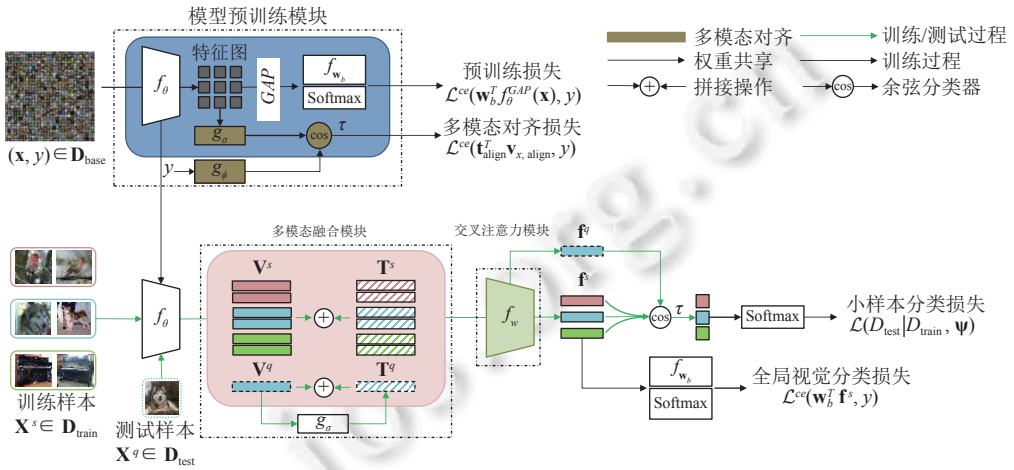

图 2 多模态引导的局部特征选择模型框架图

3.1 模型预训练过程

早期的小样本学习方法在进行小样本分类时一般使用随机初始化的模型, 在给定大量小样本训练任务 $\mathbf{D}_{\text{train}}^{\text{meta}}=\{(\mathbf{D}_{\text{train}}^{i},\mathbf{D}_{\text{test}}^{i})\}_{i=1}^{N_T}$ 的情况下, 方法按照如第 2.2 节所描述的元学习训练方式进行模型优化, 从这些小样本任务中模型将学习和搜索得到一个通用的特征空间. 但由于元学习主要强调模型对同类小样本任务的适应能力, 在上述过程中模型的特征表示能力得不到充分优化, 而在分类任务中模型特征表示能力的重要性不言而喻. 因此, 本文在进行小样本元学习训练之前先对模型进行常规分类学习预训练, 以提升模型的特征表示能力.

具体地, 设已知类别数据集 $\mathbf{D}_{\text{base}}$ 包含来自 $|\mathbf{C}_{\text{novel}}|$ 个类别的大量样本, 考虑模型由特征提取器 $f_\theta(\cdot)$ 和分类器 $f_{\mathbf{w}_b}(\cdot)$ 两部分组成, 模型预训练采用如下所示的传统模型优化方法训练模型来识别这些已知类别:

$$\boldsymbol{\Theta}\leftarrow\arg\min_{\boldsymbol{\Theta}}\mathbb{E}_{(\mathbf{x},y)\in\mathbf{D}_{\text{base}}}[\mathcal{L}^{ce}(\mathbf{w}_b^T f_\theta^{GAP}(\mathbf{x}),y)]+\mathcal{R}(\boldsymbol{\Theta}) \tag{2}$$

其中, $\boldsymbol{\Theta}=(\theta,\mathbf{w}_b)$, θ 和 $\mathbf{w}_b$ 分别代表特征提取器参数和分类器参数, y 为训练样本 $\mathbf{x}$ 的真实类别标签, $\mathcal{L}^{ce}$ 为交叉熵损失函数, $\mathcal{R}(\boldsymbol{\Theta})$ 代表正则化项. 该过程中, 通过对特征提取器输出的特征图 $f_\theta(\mathbf{x})\in\mathbb{R}^{c\times h\times w}$ 应用全局池化得到特征输出 $f_\theta^{GAP}(\mathbf{x})$, 即 $f_\theta^{GAP}(\mathbf{x})=GAP(f_\theta(\mathbf{x}))$; 在此基础上, 分类器负责将样本划分到正确的类别, 其分类误差被用来驱动本阶段的模型更新. $\mathcal{L}^{ce}$ 的具体表达式如下:

$$\mathcal{L}^{ce}=-\log(p_y(\mathbf{x})) \tag{3}$$

其中, $p_y(\mathbf{x})$ 代表 $\mathbf{w}_b^T f_\theta^{GAP}(\mathbf{x})$ 经过 Softmax 分类后属于类别 y 的概率预测值.

3.2 融合多模态信息的小样本学习过程

预训练的模型能够根据少量标签数据对若干新类别进行分类[50], 但其分类效果和经元学习优化的主流模型之间仍有一定差距, 在元学习框架下对预训练模型进行进一步的模型微调十分必要. 为此, 本文在预训练模型的基础上利用小样本训练任务 $\mathbf{D}_{\text{train}}^{\text{meta}}$ 对模型进行元学习, 如公式 (1) 所示. 对于任意 N 类 K 样本训练任务 $\mathcal{T}_i\in\mathbf{D}_{\text{train}}^{\text{meta}}$, 即 $\mathcal{T}_i=(\mathbf{D}_{\text{train}}^{i},\mathbf{D}_{\text{test}}^{i})$, 本文提出基于多模态信息引导的局部特征选择策略, 从样本局部特征中选择有代表性的局部特

征构建样本特征表示, 并基于获得的样本特征利用余弦分类器进行小样本类别预测. 为便于理解, 下面先对余弦分类器进行介绍, 然后再介绍所提出的基于多模态引导的局部特征选择策略.

3.2.1 余弦分类器

传统的图像分类任务中 (如第 3.1 节), 分类器主要采用有参线性模型, 从大量训练样本中学习能够将目标类别区分开来的分类超平面. 在小样本情况下, 对这种有参线性分类器进行训练比较困难, 因此现有小样本学习方法多采用基于特征比较的无参分类器进行小样本分类, 常用的度量函数包括欧氏距离、余弦相似度、马氏距离等. 其中, 基于余弦相似度度量的余弦分类器 (cosine classifier) 被证明能够较好地适用于小样本学习[59,60].

对于任意给定 N 类 K 样本训练任务 $\mathcal{T}_i=(\mathbf{D}^i_{\text{train}},\mathbf{D}^i_{\text{test}})$, 余弦分类器基于测试样本和训练样本之间的余弦相似关系将测试样本归类, 如公式 (4) 所示:

$$c^q=\arg\max_{c\in(1,2,\ldots,N)} p_c(\mathbf{x}^q) \tag{4}$$

其中, $p_c(\mathbf{x}^q)$ 为测试样本 $\mathbf{x}^q\in\mathbf{D}^i_{\text{test}}$ 属于类别 c 的概率预测值, 定义为:

$$p_c(x^q)=\frac{\exp(\bar{f}_c^{T}\bar{f}^q/\tau)}{\exp(\bar{f}_c^{T}\bar{f}^q/\tau)+\sum_{j\in(1,2,\ldots,N),j\neq c}\exp(\bar{f}_j^{T}\bar{f}^q/\tau)} \tag{5}$$

其中, $\bar{\mathbf{f}}^q$ 代表归一化的测试样本特征表示, $\bar{\mathbf{f}}_c$、$\bar{\mathbf{f}}_j$ 代表归一化的类别中心, τ 为尺度变换超参数. 类别中心一般通过计算同类训练样本的特征均值求得. 模型在小样本任务上的分类误差 (交叉熵损失) 将被用来驱动小样本学习过程, 误差计算如公式 (6) 所示:

$$\mathcal{L}(\mathbf{D}^i_{\text{test}}\mid\mathbf{D}^i_{\text{train}},\boldsymbol{\Psi})=-\frac{1}{N_q}\sum_{\mathbf{x}^q\in\mathbf{D}^i_{\text{test}}}\log p_y(\mathbf{x}^q) \tag{6}$$

其中, y 为测试样本 $\mathbf{x}^q$ 的真实类别标签, $p_y(\mathbf{x}^q)$ 类似于 $p_c(\mathbf{x}^q)$ 代表测试样本属于类别 y 的概率预测值.

3.2.2 多模态引导的局部特征选择

显而易见, 对于第 3.2.1 节介绍的余弦分类器, 样本的特征质量直接决定分类器的分类效果. Chen 等人[50]表明直接对预训练模型输出的样本特征图应用全局池化得到的样本特征能够比较好地应用于余弦分类器. 为得到更好的样本特征进行小样本分类, 本节介绍一种基于局部特征选择的特征提取方法. 给定小样本训练任务 $\mathcal{T}_i=(\mathbf{D}^i_{\text{train}},\mathbf{D}^i_{\text{test}})$, 对于训练样本 $\mathbf{x}^s\in\mathbf{D}^i_{\text{train}}$ 和测试样本 $\mathbf{x}^q\in\mathbf{D}^i_{\text{test}}$, 如图 2 所示, 这种基于局部特征选择的特征提取方法主要由多模态融合模块和交叉注意力模块两部分组成. 多模态融合模块结合样本的视觉特征和文本特征得到样本的多模态特征表示, 交叉注意力模块在此基础上估计样本间任意两局部特征之间的相关性, 并通过求局部特征的加权和得到最终的样本特征, 用于小样本分类.

● 多模态融合模块. 以训练样本 $\mathbf{x}^s$ 为例, 设其视觉特征和文本特征分别为 $\mathbf{V}^s=f_\theta(\mathbf{x}^s)\in\mathbb{R}^{c\times h\times w}$ 和 $\mathbf{T}^s\in\mathbb{R}^{t\times h\times w}$, 多模态融合模块将两种模态的特征拼接在一起构建样本的多模态特征:

$$\mathbf{F}^s=[\mathbf{V}^s,\mathbf{T}^s] \tag{7}$$

其中, $[\cdot]$ 为拼接运算符, 特征 $\mathbf{F}^s\in\mathbb{R}^{d\times h\times w}$, 局部特征维度 $d=c+t$. 公式 (7) 中, 样本的视觉特征 $\mathbf{V}^s$ 由特征提取器输出, 其文本特征通过将其类别标签所对应的 Word2Vec 语义词向量映射到一个对齐的语义空间中获得. 对于无标签测试样本, 由于样本类别标签的不可知性, 其文本特征需要通过将其视觉特征映射到对齐的语义空间中获得, 即 $\mathbf{T}^q=g_\sigma(\mathbf{V}^q)$, $g_\sigma(\cdot)$ 为映射函数.

● 交叉注意力模块. 给定训练样本的多模态特征 $\mathbf{F}^s\in\mathbb{R}^{d\times h\times w}$ 和测试样本的多模态特征 $\mathbf{F}^q\in\mathbb{R}^{d\times h\times w}$, 由于 $\mathbf{F}^s$ 和 $\mathbf{F}^q$ 中存在样本的文本特征, 使得样本局部特征之间具有更高的相关性, 有利于交叉注意力模块对局部特征的筛选. 两样本局部特征 $\mathbf{F}^s\in\mathbb{R}^{d\times h\times w}$ 和 $\mathbf{F}^q\in\mathbb{R}^{d\times h\times w}$ 之间的相关性定义如下:

$$r_{ij}=\left(\frac{\mathbf{F}^s_i}{\left\|\mathbf{F}^s_i\right\|_2}\right)^{T}\left(\frac{\mathbf{F}^q_j}{\left\|\mathbf{F}^q_j\right\|_2}\right) \tag{8}$$

不妨令 $\mathbf{r}^s_i\in\mathbb{R}^m$ 代表训练样本的第 i 个局部特征与测试样本所有局部特征之间的相关性, 易得训练样本相对于

测试样本的整体相关性矩阵 $\mathbf{r}^s$, 其中,

$$\mathbf{r}^s=(\mathbf{r}_1^s,\mathbf{r}_2^s,\ldots,\mathbf{r}_m^s),\ m=h\times w \tag{9}$$

相应地, 测试样本相对于训练样本的整体相关性矩阵 $\mathbf{r}^q=(\mathbf{r}^s)^T$. 类似 Hou 等人[51], 本文对相关性矩阵应用卷积核大小为 $m\times 1\times 1$ 的卷积操作 $f_w(\cdot)$ 得到每一个局部特征的相关系数, 并利用 Softmax 进行归一化处理, 如下:

$$\mathbf{A}^s=(A_1^s,A_2^s,\ldots,A_m^s) \tag{10}$$

其中,

$$A_i^S=\frac{\exp(\mathbf{w}^T\mathbf{r}_i^s/\tau)}{\exp(\mathbf{w}^T\mathbf{r}_i^s/\tau)+\sum_{j\in(1,2,\ldots,m),j\neq i}\exp(\mathbf{w}^T\cdot\mathbf{r}_j^s/\tau)} \tag{11}$$

相关系数 $\mathbf{A}^q$ 计算过程相同. 相关系数矩阵 $\mathbf{A}^s$ 和 $\mathbf{A}^q$ 实际上凸显了训练样本和测试样本之间相关的区域, 使得通过公式 (12) 获得的样本特征 $\mathbf{f}^s$ 和 $\mathbf{f}^q$ 之间对比更加鲜明.

$$\begin{aligned}\mathbf{f}^s&=GAP((1+\mathbf{A}^s)\odot\mathbf{V}^s)\\ \mathbf{f}^q&=GAP((1+\mathbf{A}^q)\odot\mathbf{V}^q)\end{aligned} \tag{12}$$

本文实验发现在公式 (12) 中使用相关系数矩阵对视觉特征进行加权得到的样本特征更有利于小样本学习.

3.3 多模态信息对齐策略

在多模态融合模块中, 样本文本特征的获取是一个比较重要的步骤. 对于类别标签已知的训练样本, 我们可以从其类别标签的语义词向量 (如 Word2Vec) 获得其文本特征表示, 但对于类别标签未知的测试样本该如何确定其文本特征并不明确. 针对这个问题, 本文提出在一个共同的空间中对视觉特征和文本特征进行语义对齐, 一旦充分训练该对齐的语义空间将能够为测试样本提供文本特征表示.

具体地, 为了在上述共同的空间进行视觉特征和文本特征的语义对齐, 本文在第 3.1 节的模型预训练阶段额外引入多模态对齐训练过程. 对于任意训练样本 $(\mathbf{x},y)\in\mathbf{D}_{\text{base}}$, 其视为成对的训练数据, 首先经视觉特征提取和文本特征提取得到样本的视觉特征 $f_\theta^{GAP}(\mathbf{x})$ 和语义词向量 $Word2Vec(y)$; 后分别将视觉特征和语义词向量映射到该共同的空间中, 如下:

$$\begin{aligned}\mathbf{v}_{\mathbf{x},\text{aligned}}&=g_\sigma(f_\theta^{GAP}(\mathbf{x}))\\ \mathbf{t}_{y,\text{aligned}}&=g_\phi(Word2Vec(y))\end{aligned} \tag{13}$$

并通过公式 (14) 约束它们之间的对齐关系:

$$\mathbf{\Phi}\leftarrow\arg\min_{\mathbf{\Phi}}\mathbb{E}_{(\mathbf{x},y)\in\mathbf{D}_{\text{base}}}[\mathcal{L}^{ce}(\mathbf{t}_{\text{aligned}}^T\mathbf{v}_{\mathbf{x},\text{aligned}},y)] \tag{14}$$

其中, $\mathbf{\Phi}=(\sigma,\phi)$ 为空间变换参数, 映射函数 $g_\sigma(\cdot)$, $g_\phi(\cdot)$ 为 1×1 卷积操作, $\mathbf{t}_{\text{aligned}}$ 为所有已知类别 $\mathbf{C}_{\text{base}}$ 在空间中的文本特征表示. 可以看到, 公式 (14) 将视觉特征和文本特征对齐转换成分类问题. 该过程中, 我们使用了类似第 3.2.1 节介绍的余弦分类器. 基于这种语义对齐, 小样本学习时对于无标签测试样本我们可以很方便地基于其样本的视觉生成其文本特征, 见第 3.2.2 节.

4 实验分析

4.1 实验数据集

本文在 MiniImageNet, CIFAR-FS 和 FC-100 这 3 个数据集上对方法的有效性进行验证. 其中:

● MiniImageNet 数据集[3]是 ImageNet 数据集[61]的一个子集, 包含 100 个类别, 每个类别包含 600 张图像, 每幅图像的尺寸为 84×84. 100 个类别被划分为 64 类用于模型训练、16 类用于模型验证和超参数选择、20 类用于模型测试.

● CIFAR-FS 数据集[8]是基于 CIFAR-100[62]构建的小样本数据集, 它包含 100 个类别, 每个类别包含 600 张图像, 每幅图像的尺寸为 32×32. 数据集被划分为 64 类用于模型训练、16 类用于模型验证和超参数选择、20 类用于模型测试.

● FC-100 数据集[63]也是基于 CIFAR-100[62]构建的小样本数据集, 它将 CIFAR-100 的 100 个类别划分为 20

个超类, 其中 12 个超类 (60 个类别) 用于模型训练, 4 个超类 (20 个类别) 用于模型验证和超参数选择, 4 个超类 (20 个类别) 用于模型测试. 数据集中每幅图像的尺寸和 CIFAR-FS 数据集相同, 为 32×32.

4.2 实验设置

本文类似 Chen 等人[50]采用 ResNet-12 作为主干网络进行小样本学习. ResNet-12 网络由 4 个残差模块组成, 每个残差模块有 3 个卷积层, 每个卷积层由一个 3×3 的卷积核, 一个 BatchNorm 归一化层以及一个 ReLU 激活函数组成. 4 个残差块的卷积层通道数分别为 64、128、256、512, 在每个残差块后均使用了 2×2 的最大池化层对样本特征进行下采样. 每个输入样本最终将被表示为一个 512 维的特征向量.

模型预训练在各数据集的训练集分支上对网络进行大规模图像分类学习, 学习率初始设置为 0.1, 优化器为 SGD, 动量设置为 0.9. 在 MiniImageNet、CIFAR-FS 和 FC-100 数据集上, 训练批次大小 (batch size) 设置为 128, 训练迭代 100 次, 学习率在第 90 次训练时衰减一次, 衰减因子为 0.1. 在模态对齐部分, 对数据集类别标签应用 Word2Vec 提取得到维度为 300 的语义词向量, 余弦分类器使用自适应尺度变换超参数 τ, 初始值设置为 10.

在融合多模态信息的小样本学习阶段, 利用预训练的 ResNet-12 网络初始化特征提取器和全局视觉分类器部分. 全局视觉分类器一经初始化便进行固化处理, 在小样本学习过程中将不进行优化. 模型在从训练集分支上随机采样的 72 000 个小样本任务上进行训练, 训练迭代 60 次, 每次 1 200 个任务. 每次训练迭代结束后, 在从验证集分支上随机采样的 2 000 个小样本任务上对模型性能进行评估, 并最终在从测试集分支上随机采样的 2 000 个小样本任务上对模型泛化性能进行验证. 模型训练采用 SGD 优化器, 动量大小设置为 0.9. 初始学习率设置为 0.1. 该阶段学习率分别在第 30 次迭代、40 次迭代和第 50 次迭代时将学习速率下降到 0.006、0.001 2 和 0.000 24. 小样本余弦分类器同样使用自适应尺度变换超参数 τ, 初始值设置为 10.

表 1 MiniImageNet 数据集上的对比实验结果 (%)

方法	主干网络	MiniImageNet	
		5类单样本	5类5样本
MAML[6]	ConvNet-4	48.70 ± 1.84	63.11 ± 0.92
MatchingNet[3]	ConvNet-4	43.56 ± 0.84	55.31 ± 0.73
ProtoNet[5]	ConvNet-4	49.42 ± 0.78	68.20 ± 0.66
RelationNet[56]	ConvNet-4	50.44 ± 0.82	65.32 ± 0.70
R2D2[8]	ConvNet-4	51.20 ± 0.60	68.80 ± 0.10
AdaResNet[64]	ResNet-12	56.88 ± 0.62	71.94 ± 0.57
TADAM[63]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30
TEWAM[14]	ResNet-12	60.07 ± n/a	75.90 ± n/a
MTL[48]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80
Variational FSL[67]	ResNet-12	61.23 ± 0.26	77.69 ± 0.17
DSN[65]	ResNet-12	62.64 ± 0.66	78.83 ± 0.45
FBM+MTL[66]	ResNet-12	61.41 ± 1.87	76.11 ± 0.92
ModelRegression[68]	ResNet-12 (D)	61.94 ± 0.20	76.24 ± 0.14
RFS[23]	ResNet-12 (D)	62.02 ± 0.63	79.64 ± 0.44
MetaOptNet[9]	ResNet-12 (D)	62.64 ± 0.20	78.63 ± 0.14
Meta-Baseline[50]	ResNet-12	63.17 ± 0.23	79.26 ± 0.17
CAN[51]	ResNet-12	63.85 ± n/a	79.44 ± n/a
本文	ResNet-12	**66.17 ± 0.20**	**81.84 ± 0.15**

4.3 主要实验结果

4.3.1 MiniImageNet 数据集上的结果

本文首先在 MiniImageNet 数据集上, 与相关的小样本图像分类技术进行性能比较, 如表 1 所示. 文献 [3,5,6,8,56]

使用 4 层卷积网络作为主干网络; 文献 [14,48,50,51,63−67] 使用 ResNet-12 作为主干网络; 文献 [9,23,68] 的主干网络在 ResNet-12 基础上, 参考文献 [69] 使用 Dropblock 作为正则化项, 并将 4 个残差块的卷积层通道数从 (64、128、256、512) 更改为 (64、160、320、640). 实验分别在 5 类单样本 (5-way 1-shot) 和 5 类 5 样本 (5-way 5-shot) 两种小样本任务设置下进行, 并报告了一个置信区间为 95% 的平均小样本分类准确率 (%), 没有说明置信区间的对比方法使用“n/a”表示. 可以观察到, 在 MiniImageNet 数据集上, 本文的方法优于所有的对比方法. 更具体地说, 本文分别在 5 类单样本和 5 类 5 样本下实现了 66.13% 和 81.80% 的平均准确性 (Avg). 与当时最领先的对比方法 (CAN) 相比, 本文的方法 5 类单样本/5 类 5 样本的 Avg 分别比该方法高 2.32%/2.40%.

4.3.2 CIFAR-FS 数据集上的结果

在 CIFAR-FS 数据集上的实验结果如表 2 所示. 文献 [6,8,56] 使用 4 层卷积网络作为主干网络; 文献 [5,14,65] 使用 ResNet-12 作为主干网络; 文献 [9,23,70] 使用 ResNet-12(D) 作为主干网络. 实验与 MiniImageNet 使用相同的小样本任务设置, 并报告了方法在该数据集上置信区间为 95% 的平均小样本分类准确率 (%), 没有说明置信区间的对比方法使用“n/a”表示. 可以看到, 在 CIFAR-FS 数据集上, 本文方法分别在 5 类单样本和 5 类 5 样本下实现了 74.30% 和 85.21% 的平均准确性 (Avg). 与当时最领先的对比方法 (DSN) 相比, 本文的方法在 5 类单样本下的 Avg 比该方法高 2.0%, 在 5 类 5 样本下的 Avg 比该方法高 0.11%.

表 2 CIFAR-FS 数据集上的对比实验结果 (%)

方法	主干网络	CIFAR-FS	
		5类单样本	5类5样本
MAML[6]	ConvNet-4	58.90 ± 1.90	71.50 ± 1.00
RelationNet[56]	ConvNet-4	55.00 ± 1.00	69.30 ± 0.80
R2D2[8]	ConvNet-4	65.30 ± 0.20	79.40 ± 0.10
ShotFree[70]	ResNet-12 (D)	69.20 ± n/a	84.70 ± n/a
TEWAM[14]	ResNet-12	70.40 ± n/a	81.30 ± n/a
RFS[23]	ResNet-12 (D)	71.50 ± 0.80	86.00 ± 0.50
MetaOptNet[9]	ResNet-12 (D)	72.00 ± 0.70	84.20 ± 0.50
ProtoNet[5]	ResNet-12	72.20 ± 0.70	83.50 ± 0.50
DSN[65]	ResNet-12	72.30 ± 0.80	85.10 ± 0.60
本文	ResNet-12	**74.30± 0.50**	**85.21± 0.30**

4.3.3 FC-100 数据集上的结果

类似地, 方法在 FC-100 数据集上的实验结果如表 3 所示. 文献 [5,63] 使用 ResNet-12 作为主干网络; 文献 [9] 使用 ResNet-12 (D) 作为主干网络. 实验与 MiniImageNet 使用相同的小样本任务设置, 并报告了方法在 FC-100 数据集上置信区间为 95% 的平均小样本分类准确率 (%). 实验结果表明, 在 FC-100 数据集上本文分别在 5 类单样本和 5 类 5 样本设置下实现了 42.42% 和 57.24% 的平均准确性 (Avg). 与当时最领先的对比方法 (MetaOptNet) 相比, 本文的方法在 5 类单样本上的 Avg 比该方法高 1.32%, 在 5 类 5 样本上的 Avg 比该方法高 1.74%. 在 MiniImageNet, CIFAR-FS 和 FC-100 这 3 个数据集上的对比实验证明了所提出方法的有效性.

表 3 FC-100 数据集上的对比实验结果 (%)

方法	主干网络	FC-100	
		5类单样本	5类5样本
TADAM[63]	ResNet-12	40.10 ± 0.40	56.10 ± 0.40
ProtoNet[5]	ResNet-12	37.50 ± 0.60	52.50 ± 0.60
MetaOptNet[9]	ResNet-12 (D)	41.10 ± 0.60	55.50 ± 0.60
本文	ResNet-12	**42.42 ± 0.50**	**57.24 ± 0.50**

4.4 消融实验

4.4.1 模型预训练技术的有效性验证

为了验证模型预训练技术的有效性, 本文删除多模态融合模块, 排除由于该模块对实验结果的影响, 通过是否加载预训练模型来观察所提出的方法在 MiniImageNet 数据集上的性能变化 (%). 随机初始化模型的方法记做“无 P 无 M” (无 Pre-train 无 Multi), 加载预训练模型的方法记做“有 P 无 M”. 实验结果见后文表 4.

如表 4 所示, 在两种小样本任务设置下, 加载预训练模型的方法比随机初始化模型的方法在平均准确性 (Avg) 上分别产生了 1.77% 和 1.58% 的性能增益. 由此可见模型预训练技术对提高模型特征表达能力有着明显作用, 通过预训练, 可以获得更高的小样本学习精度.

4.4.2 多模态学习技术的有效性验证

为了进一步验证多模态学习技术的有效性, 同时判断使用模型预训练技术与多模态学习技术的方法是否能够获得最佳实验效果, 本文加载预训练模型, 通过是否使用多模态融合模块来观察所提出的方法在 MiniImageNet 数据集上的性能变化 (%), 删除多模态融合模块的方法记做“有 P 无 M”, 使用多模态融合模块的方法记做“有 P 有 M”. 实验结果见表 4.

如表 4 所示, 在两种小样本任务设置下, 使用多模态融合模块的方法相比没有使用多模态融合模块的方法在平均准确性 (Avg) 上分别产生了 0.55% 和 0.82% 的性能增益. 这证明了使用多模态融合模块能够提升样本特征的表达能力, 也证明了同时使用模型预训练技术与多模态学习技术的实验方法能够获得最佳实验效果.

4.4.3 不同特征对小样本分类结果的影响

大多数的小样本学习任务均使用样本的视觉特征进行图像分类, 本文也不例外. 而本文中除视觉特征外, 还涉及文本特征和多模态特征, 为了验证其他特征对小样本学习是否适用, 本文使用多模态相关系数矩阵分别对文本特征、视觉特征以及这两种特征融合后的多模态特征进行加权, 得到了 3 种不同的样本特征. 本文在“有 P 有 M”实验设置下, 观察采用 3 种不同样本特征时方法在 MiniImageNet 数据集上的性能变化 (%), 实验结果见表 5. 其中, 我们将使用文本特征的方法记做“Txt 特征”, 使用视觉特征的方法记做“Visual 特征”, 使用多模态融合特征的方法记做“Multi 特征”.

表 4 模型预训练技术与多模态学习技术的有效性验证 (%)

方法	5类单样本	5类5样本
无P无M	63.85	79.44
有P无M	65.62	81.02
有P有M	**66.17**	**81.84**

表 5 在 MiniImageNet 数据集上 3 种样本特征的对比实验结果 (%)

方法	5类单样本	5类5样本
Txt特征	63.44	79.96
Multi特征	65.76	80.35
Visual特征	**66.17**	**81.84**

如表 5 所示, 在两种小样本任务设置下, 本文的方法拥有比另外两种特征更高的平均准确性 (Avg). 更具体地说, 使用视觉特征的方法在 5 类单样本下的 Avg 比使用文本特征和多模态特征的方法高 2.73% 和 0.41%, 在 5 类 5 样本下的 Avg 比使用文本特征和多模态特征的方法高 1.88% 和 1.49%. 这证明了对视觉特征进行加权得到的样本特征能够更好地适应小样本分类任务.

5 总 结

针对现有的元学习方法识别特征单一、特征表达不充分、不准确等问题, 本文提出多模态引导的局部特征选择小样本学习方法, 并对预训练技术和多模态学习技术进行了研究. 该方法首先在大量已知类别数据上对模型进行预训练; 然后在元学习阶段, 将样本的视觉特征和文本特征进行融合并进行局部特征选择; 最后对选择后的样本特征进行小样本学习. 通过在 MiniImageNet、CIFAR-FS 和 FC-100 这 3 个基准数据集上的对比实验结果, 可以有效地验证提高样本特征质量能够提升模型的小样本学习精度. 与此同时, 本文分别验证了预训练技术和多模态学

习技术的有效性, 证明了本文所提出的学习方法可以有效提高模型的特征表达能力.

References:

[1] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Advances in Neural Information Processing Systems. Curran Associates Inc., 2017.

[2] Henaff O. Data-efficient image recognition with contrastive predictive coding. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 4182–4192.

[3] Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proc. of the 30th Conf. on Neural Information Processing Systems. Barcelona: ACM, 2016. 3637–3645.

[4] Pahde F, Puscas M, Klein T, Nabi M. Multimodal prototypical networks for few-shot learning. In: Proc. of the 2021 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 2644–2653. [doi: 10.1109/WACV48630.2021.00269]

[5] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 4080–4090.

[6] 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.

[7] Li ZG, Zhou FW, Chen F, Li H. Meta-SGD: Learning to learn quickly for few-shot learning. arXiv:1707.09835, 2017.

[8] Bertinetto L, Henriques JF, Torr PHS, Vedaldi A. Meta-learning with differentiable closed-form solvers. arXiv:1805.08136, 2018.

[9] Lee K, Maji S, Ravichandran A, Soatto S. Meta-learning with differentiable convex optimization. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10657–10665. [doi: 10.1109/CVPR.2019.01091]

[10] Hariharan B, Girshick R. Low-shot visual recognition by shrinking and hallucinating features. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3018–3027. [doi: 10.1109/ICCV.2017.328]

[11] Zhang HG, Zhang J, Koniusz P. Few-shot learning via saliency-guided hallucination of samples. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2770–2779. [doi: 10.1109/CVPR.2019.00288]

[12] Luo QX, Wang LF, Lv JG, Xiang SM, Pan CH. Few-shot learning via feature hallucination with variational inference. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3963–3972. [doi: 10.1109/WACV48630.2021.00401]

[13] Li K, Zhang YL, Li KP, Fu Y. Adversarial feature hallucination networks for few-shot learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13470–13479. [doi: 10.1109/CVPR42600.2020.01348]

[14] Qiao LM, Shi YM, Li J, Tian YH, Huang TJ, Wang YW. Transductive episodic-wise adaptive metric for few-shot learning. In: Proc. of the 2019 IEEE Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3603–3612. [doi: 10.1109/ICCV.2019.00370]

[15] Qi GD, Yu HM, Lu ZH, Li SZ. Transductive few-shot classification on the oblique manifold. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 8392–8402. [doi: 10.1109/ICCV48922.2021.00830]

[16] Liu YB, Lee J, Park M, Kim S, Yang E, Hwang SJ, Yang Y. Learning to propagate labels: Transductive propagation network for few-shot learning. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–14.

[17] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: PMLR, 2013. 1139–1147.

[18] Zhang HY, Cissé M, Dauphin YN, Lopez-Paz D. Mixup: Beyond empirical risk minimization. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–13.

[19] DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552, 2017.

[20] Yun S, Han D, Chun S, Oh SJ, Yoo Y, Choe J. CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6023–6032. [doi: 10.1109/ICCV.2019.00612]

[21] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.

[22] Raghu A, Raghu M, Bengio S, Vinyals O. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–21.

[23] Tian YL, Wang Y, Krishnan D, Tenenbaum JB, Isola P. Rethinking few-shot image classification: A good embedding is all you need? In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 266–282. [doi: 10.1007/978-3-030-58568-6_16]

[24] He J, Hong RC, Liu XL, Xu ML, Sun QR. Revisiting local descriptor for improved few-shot classification. ACM Trans. on Multimedia Computing, Communications, and Applications, 2022, 18(2s): 127. [doi: 10.1145/3511917]

[25] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern

Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

[26] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

[27] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 10012–10022. [doi: 10.1109/ICCV48922.2021.00986]

[28] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 1–14.

[29] Wu ZY, Li YW, Guo LH, Jia K. PARN: Position-aware relation networks for few-shot learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6659–6667. [doi: 10.1109/ICCV.2019.00676]

[30] Doersch C, Gupta A, Zisserman A. CrossTransformers: Spatially-aware few-shot transfer. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 21981–21993.

[31] Tsai YHH, Bai SJ, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]

[32] Hong DF, Gao LR, Yokoya N, Yao J, Chanussot J, Du Q, Zhang B. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. on Geoscience and Remote Sensing, 2021, 59(5): 4340–4354. [doi: 10.1109/TGRS.2020.3016820]

[33] Wang YK, Huang WB, Sun FC, Xu TY, Rong Y, Huang JZ. Deep multimodal fusion by channel exchanging. In: Proc. of the 34th Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2020. 4835–4845.

[34] Li FF, Fergus R, Perona P. A Bayesian approach to unsupervised one-shot learning of object categories. In: Proc. of the 9th IEEE Int'l Conf. on Computer Vision. Nice: IEEE, 2003. 1134–1141. [doi: 10.1109/ICCV.2003.1238476]

[35] Xu WJ, Xian YQ, Wang JN, Schiele B, Akata Z. Attribute prototype network for zero-shot learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 21969–21980.

[36] Hubert Tsai YH, Huang LK, Salakhutdinov R. Learning robust visual-semantic embeddings. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3571–3580. [doi: 10.1109/ICCV.2017.386]

[37] Xing C, Rostamzadeh N, Oreshkin BN, Pinheiro PO. Adaptive cross-modal few-shot learning. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 32.

[38] Zhu YH, Min WQ, Jiang SQ. Attribute-guided feature learning for few-shot image recognition. IEEE Trans. on Multimedia, 2021, 23: 1200–1209. [doi: 10.1109/tmm.2020.2993952]

[39] Sulc M, Picek L, Matas J, Jeppesen TS, Heilmann-Clausen J. Fungi recognition: A practical use case. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision. Snowmass: IEEE, 2020. 2316–2324. [doi: 10.1109/WACV45572.2020.9093624]

[40] Kiss N, Czùni L. Mushroom image classification with CNNs: A case-study of different learning strategies. In: Proc. of the 12th Int'l Symp. on Image and Signal Processing and Analysis. Zagreb: IEEE, 2021. 165–170. [doi: 10.1109/ISPA52656.2021.9552053]

[41] Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10781–10790. [doi: 10.1109/CVPR42600.2020.01079]

[42] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[43] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[44] Zhu JH, Xia YC, Wu LJ, He D, Qin T, Zhou WG, Li HQ, Liu TY. Incorporating BERT into neural machine translation. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–18.

[45] Garg S, Ramakrishnan G. BAE: BERT-based adversarial examples for text classification. arXiv:2004.01970, 2020.

[46] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 1597–1607.

[47] He KM, Fan HQ, Wu YX, Xie SN, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9729–9738. [doi: 10.1109/CVPR42600.2020.00975]

[48] Sun QR, Liu YY, Chua TS, Schiele B. Meta-transfer learning for few-shot learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 403–412. [doi: 10.1109/CVPR.2019.00049]

[49] Chen WY, Liu YC, Kira Z, Wang YCF, Huang JB. A closer look at few-shot classification. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–16.

[50] Chen YB, Wang XL, Liu Z, Xu HJ, Darrell T. A new meta-baseline for few-shot learning. arXiv:2003.04390v2, 2020.

[51] 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: ACM, 2019. 4003–4014.

[52] Huang ST, Zhang M, Kang YC, Wang DL. Attributes-guided and pure-visual attention alignment for few-shot recognition. Proc. of the 2021 AAAI Conf. on Artificial Intelligence, 2021, 35(9): 7840–7847. [doi: 10.1609/aaai.v35i9.16957]

[53] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR, 2015. 1–30.

[54] Hoffman J, Tzeng E, Donahue J, Jia YQ, Saenko K, Darrell T. One-shot adaptation of supervised deep convolutional models. arXiv: 1312.6204, 2013.

[55] Santoro A, Bartunov S, Botvinick MM, Wierstra D, Lillicrap TP. Meta-learning with memory-augmented neural networks. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York City: PMLR, 2016. 1842–1850.

[56] Sung F, Yang YX, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to compare: Relation network for few-shot learning. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1199–1208. [doi: 10.1109/CVPR.2018.00131]

[57] Schmidhuber J. Evolutionary principles in self-referential learning [Ph.D. Thesis]. München: Technische Universität München, 1987.

[58] Hinton GE, Plaut DC. Using fast weights to deblur old memories. In: Proc. of the 9th Annual Conf. of the Cognitive Science Society. 1987. 177–186.

[59] Qi H, Brown M, Lowe DG. Low-shot learning with imprinted weights. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5822–5830. [doi: 10.1109/CVPR.2018.00610]

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

[61] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Huang ZH, Ma S, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. Imagenet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

[62] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Handbook of Systemic Autoimmune Diseases, 2009: 1–60.

[63] Oreshkin BN, Rodríguez P, Lacoste A. TADAM: Task dependent adaptive metric for improved few-shot learning. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 719–729.

[64] Munkhdalai T, Yuan X, Mehri S, Trischler A. Rapid adaptation with conditionally shifted neurons. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 3661–3670.

[65] Simon C, Koniusz P, Nock R, Harandi M. Adaptive subspaces for few-shot learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4136–4145. [doi: 10.1109/CVPR42600.2020.00419]

[66] Yang P, Ren SG, Zhao Y, Li P. Calibrating CNNs for few-shot meta learning. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022. 2090–2099. [doi: 10.1109/WACV51458.2022.00048]

[67] Zhang J, Zhao CL, Ni BB, Xu MH, Yang XK. Variational few-shot learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1685–1694. [doi: 10.1109/ICCV.2019.00177]

[68] Wang YX, Hebert M. Learning to learn: Model regression networks for easy small sample learning. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 616–634. [doi: 10.1007/978-3-319-46466-4_37]

[69] Ghiasi G, Lin TY, Le QV. DropBlock: A regularization method for convolutional networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 10750–10760.

[70] Ravichandran A, Bhotika R, Soatto S. Few-shot learning with embedded class models and shot-free meta training. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 331–339. [doi: 10.1109/ICCV.2019.00042]

吕天根(1997－), 男, 硕士生, CCF 学生会员, 主要研究领域为小样本学习.

何军(1992－), 男, 博士, 主要研究领域为模式识别, 小样本学习, 弱监督学习.

洪日昌(1981－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体技术, 人工智能, 大数据.

胡社教(1964－), 男, 博士, 教授, 主要研究领域为智能检测与信号处理, 智能配变终端系统, 嵌入式控制系统.