doi:10.11887/j.cn.202401002 http://journal.nudt.edu.cn

多模态交叉解耦的少样本学习方法

冀 中[1*],王思迪[1],于云龙[2]
(1. 天津大学 电气自动化与信息工程学院,天津 300072;2. 浙江大学 信息与电子工程学院,浙江 杭州 310027)

摘 要:当前的多模态少样本学习方法忽视了属性间差异对正确识别样本类别的影响。针对这一问题,提出一种利用多模态交叉解耦的方法,通过解耦不同属性语义特征,并经过特征重建学习样本的本质类别特征,缓解类别属性差异对类别判别的影响。在两个属性差异较大的基准少样本数据集 MIT-States 和 C-GQA 上进行的大量实验表明,所提方法较现有方法有较大的性能提升,充分验证了方法的有效性,表明多模态交叉解耦的少样本学习方法能够提升识别少量测试样本的分类性能。

关键词:少样本学习;多模态学习;特征解耦;属性
中图分类号:TP18 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2024)01-012-10

Multimodal cross-decoupling for few-shot learning

JI Zhong[1*], *WANG Sidi*[1], *YU Yunlong*[2]
(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: Current multi-modal few-shot learning methods overlook the impact of inter-attribute differences on accurately recognizing sample categories. To address this problem, a multimodal cross-decoupling method was proposed which could decouple semantic features with different attributes and reconstruct the essential category features of samples, aiming to alleviate the impact of category attribute differences on category discrimination. Extensive experiments on two benchmark few-shot datasets MIT-States and C-GQA with large attribute discrepancy indicates that the proposed method outperforms the existing approaches, which fully verifies its effectiveness, indicating that the multimodal cross-decoupling few-shot learning method can improve the classification performance of identifying few test samples.

Keywords: few-shot learning; multimodal learning; feature decoupling; attribute

近年来,深度学习技术在人工智能领域取得了令人瞩目的成绩,例如图像分类[1-2]、目标检测[3-4]和行人重识别[5-6]等。这些方法大多都依赖于大量的标注数据。然而,人类可以通过少量的样本学习到新知识,这种“举一反三”的能力极大地启发了研究人员,少样本学习[7-8]就是其中一类有代表性的技术,其目的是从有限数量的标注数据中快速学习类别知识。

本文针对图像少样本分类技术展开研究,现有方法通常通过有限样本学习一个可迁移模型,利用少量视觉样本识别新的类别[9-10]。然而,现有方法更多地关注样本视觉表征较为单一的类别,当样本视觉表征多样化时,较难准确分类。

具体地,在少样本学习中由于可用于训练的标注样本数量极为有限,因此这些样本较难完全反映类别的多样性视觉表征。这本质是由于一些类别具有多元属性。例如“苹果”同时具有“红色”和“黄色”属性。属性多元化会造成视觉表征复杂度的提高,从而损害属性无关视觉特征的提取,本文将这种问题称为“多元属性问题”。如图1所示,当有限标注的样本都是红色苹果时,该类别特征将会突显其“红色”视觉表征,而当测试样本是黄色苹果时,此时由于香蕉常带有“黄色”表皮属性,因此少见的“黄色”苹果易被错分为香蕉。

如何从有限样本中提取到属性无关的类别特征是解决上述问题的关键所在。在少样本学习中,视觉特征往往受到对应属性的影响,容易偏向学习表层属性而忽略类别本质。相比之下,样本的类别和属性信息易于获取,并能指导视觉特征

收稿日期:2022-06-21
基金项目:浙江省重点研发计划资助项目(2021C01119);国家自然科学基金资助项目(62176178,62002320,U19B2043)
*第一作者:冀中(1979—),男,河北保定人,教授,博士,博士生导师,E-mail:jizhong@tju.edu.cn

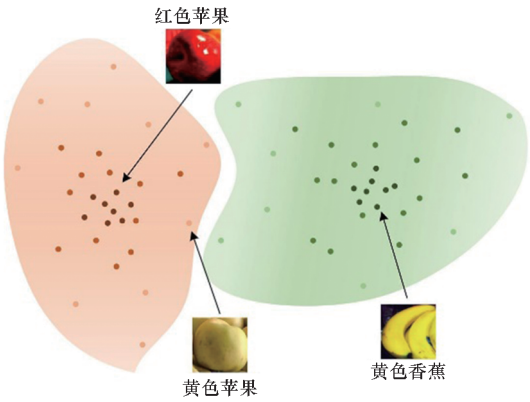


图 1　多元属性样本分布

Fig. 1　Distribution of diverse attribute samples

的提取。因此,一些少样本学习方法[11-12]借鉴多模态学习的思想,通过利用辅助模态如类别属性、文本描述等来帮助分类,在一定程度上可缓解多元属性问题的影响。例如,自适应跨模态(adaptive modality mixture mechanism, AM3)方法[11]自适应地将语义和视觉信息组合起来,通过语义中的类别信息指导视觉特征提取。属性引导注意力模块(attributes-guided attention module, AGAM)[12]使用类别语义信息引导特征提取,与视觉信息结合进行注意力对齐。然而,这些方法均侧重于使用单一的类别语义信息促进视觉特征的提取,忽视了类别中存在属性的多样性以及不同属性间的差异对正确识别样本类别作用的不同[11-12]。

针对这一问题,本文提出一种多模态交叉解耦(multimodal cross-decoupling, MCD)的少样本图像分类方法,通过学习样本的视觉特征、类别语义特征和属性语义特征,解耦出样本的多元属性,来减少属性多样化带来的分类差异,缓解多元属性问题对分类的影响。

现有的少样本学习方法大致可以分为三个类别:基于优化的方法、基于生成的方法和基于度量的方法。

基于优化的少样本学习方法通过学习一个元学习器和调整优化算法的参数来适应少样本分类任务。这些方法通常在基类样本上进行训练,得到一个学习模型,然后在新类样本的少样本任务中进行微调。与模型无关的元学习(model-agnostic meta-learning, MAML)方法[2]通过在元训练阶段让模型学习一个较好的初始化参数,并在新类少样本任务上经过几步梯度更新优化模型参数。具有潜在嵌入优化(latent embedding optimization, LEO)的元学习方法[13]改进了MAML 方法,通过编码器将特征表达映射到低维隐空间,并使用解码器将隐向量转化为高维模型参数,通过优化隐向量来进行参数更新。

基于生成的少样本学习方法,其思路是通过对样本进行变换以达到数据增强的目的,或是使用生成对抗网络来扩大样本空间,从而弥补类别样本数量少的不足。Chen 等[14]提出的图像变形元网络(image deformation meta-networks, IDeMe-Net)使用网络学习生成不同的变换图形来扩充数据,并提取特征信息。Li 等[15]提出的对抗性特征幻觉网络(adversarial feature hallucination networks, AFHN)利用生成对抗网络(generative adversarial networks, GAN)生成多样性样本,并引入分类和反崩溃正则项。

基于度量的少样本学习方法通过将支持集和查询集的特征映射到一个新的度量空间,并学习可迁移的距离度量来进行分类。原型网络(prototypical networks, PN)[7]将同类别样本特征均值作为类原型,并使用欧氏距离计算查询样本与类原型之间的距离来分类。关系网络(relation network, RN)[8]通过神经网络学习动态的度量模型,并利用该模型计算查询样本与支持样本之间的相似度。Li 等[16]提出的深度最近邻网络(deep nearest neighbor neural network, DN4)方法通过使用局部描述子来比较查询图像与支持集之间的相似度。

多模态学习研究不同模态数据的机器学习问题,通过挖掘模态间的互补性或独立性来表征多模态数据。多模态表示学习应用于编码不同模态数据中的语义信息,并学习各模态的特征与映射关系,其在跨模态检索[17-18]、图文匹配[19-20]、零样本学习[21-22]等任务中都有着较为重要的作用。

多模态少样本学习方法通过引入语义信息,联合训练文本和视觉特征,提升少样本分类任务性能。例如,AM3 方法[11]利用语义表征提供先验知识来补充视觉信息,通过凸组合将视觉和语义表征结合起来进行分类。模态交替传播网络(modal-alternating propagation network, MAP-Net)[23]利用图传播指导语义图更新,通过计算视觉特征相似性弥补缺失的语义特征。属性指导特征学习(attribute-guided feature learning, AGFL)方法[24]通过属性相关表示建立联系,增强相同属性在不同类别的表达。

本文所提方法是一种基于度量学习的多模态少样本学习方法。与以往方法不同,本文方法针对少样本多元属性问题,能够识别带有相似属性

的同类样本，同时还能够具备识别出有不同属性的该类样本的能力。

1 方法实现

1.1 问题定义

假设存在一个给定数据集 $D=\{D^{\text{train}},D^{\text{test}}\}$，其中训练集 D^{train} 和测试集 D^{test} 在样本空间中不相交。训练集样本 D^{train} 包含了大量的类别，其标签空间为 $C^{\text{base}}=\{c^1,c^2,\cdots,c^n\}$，$n$ 表示总类别数量。标准的少样本分类任务（每一个任务包含 N 个类别、每个类别包含 K 个样本）被称作 N-way K-shot 任务，一般 K 是比较小的整数，如 1 或 5。

在少样本分类任务中 K 值较小导致训练样本不足，元学习方法通过使用大量训练集数据 D^{train} 合理解决这一问题。常见的基于度量学习的少样本分类方法基于 episode 的形式实现元学习的训练与测试，并通过 episode 随机采样形成上述的 N-way K-shot 任务。通常情况下，每一个 episode 包含一个支持集 X^{S} 和一个查询集 X^{Q}。在支持集 X^{S} 中所有样本 x 都是标注数据，通过特征提取器 f 得到样本特征 $f(x)$，同时可以利用原型网络[7]计算类别原型 P 作为支持集类别的样本特征。而在查询集 X^{Q} 中每个样本都是无标注数据。

1.2 整体框架

针对多元属性问题提出的多模态交叉解耦的少样本分类方法整体框架如图 2 所示。模型方法由多模态交叉解耦和特征重建两个部分组成。在元训练阶段，多模态交叉解耦模块将样本的视觉特征、多种语义特征进行解耦，以得到解耦后的类别特征和属性特征。为保证解耦出来的特征能够准确地表示原样本，利用特征重建模块将解耦后的特征重新进行融合，并与原本的全局视觉特征进行约束。在元测试阶段，因为查询集没有语义信息，所以使用自解耦来代替交叉解耦，并与支持集样本解耦后的类别特征进行分类损失判别。

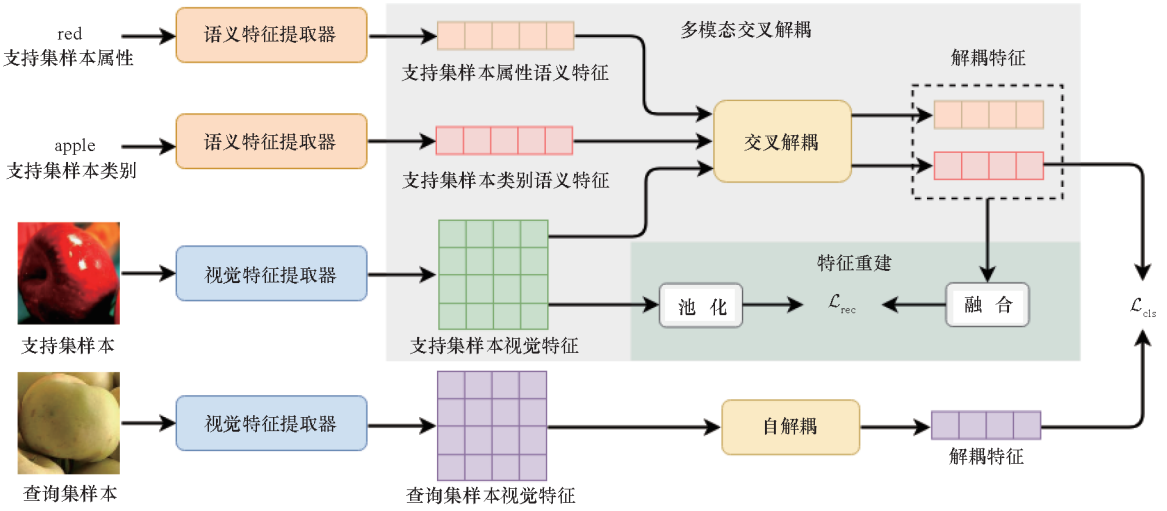


图 2 所提多模态交叉解耦的少样本分类方法框架

Fig. 2 Proposed framework of multimodal cross-decoupling few-shot classification

1.3 多模态交叉解耦

为了缓解多元属性问题，提出利用信息交叉解耦的思想，从多模态信息中学习有效的类别信息，弱化属性特征对分类效果的影响。

给定基类支持集样本 X^{S}，其类别和属性的语义信息分别表示为 X^{L} 和 X^{A}。每个输入的支持集样本 $x_i\in X^{\text{S}}$，通过视觉特征提取器 f 得到对应的局部视觉特征 $f(x_i)\in\mathbf{R}^{h\times w\times d}$，其中 h、w 分别表示特征图的高和宽，d 表示特征的维度。对于每个样本的每对语义信息，类别 x_i^{L} 和属性 x_i^{A} 分别通过语义特征提取器 g 得到对应的类别特征 $g(x_i^{\text{L}})\in\mathbf{R}^{1\times d}$ 和属性特征 $g(x_i^{\text{A}})\in\mathbf{R}^{1\times d}$。不同语义信息对同类样本视觉特征的影响作用不同，因此考虑利用样本的视觉特征对类别语义特征和多元属性语义特征进行解耦。具体来说，如图 3 所示（图中⊙表示点乘），采用交叉注意力机制，利用视觉特征与相关语义的相似性分别提取解耦后的类别特征和语义特征。

首先，采用交叉注意力机制挖掘视觉特征 $f(x_i)$ 和类别语义特征 $g(x_i^{\text{L}})$ 的关系相似性，并线性映射成三个维度为 d' 的特征，分别是查询 $\boldsymbol{Q}\in\mathbf{R}^{1\times d'}$，键 $\boldsymbol{K}\in\mathbf{R}^{hw\times d'}$ 和值 $\boldsymbol{V}\in\mathbf{R}^{hw\times d'}$，其中 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别通过式（1）、式（2）和式（3）计算得到。

$$\boldsymbol{Q}=g(x_i^{\text{L}})\cdot\boldsymbol{W}_Q^{\text{T}} \tag{1}$$

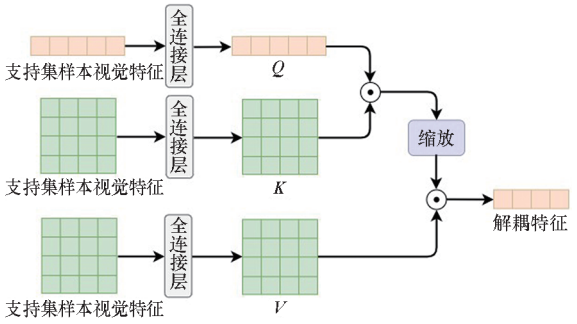


图 3　交叉解耦示意图

Fig. 3　Illustration of cross-decoupling

$$\boldsymbol{K}=f(x_i)\cdot\boldsymbol{W}_K^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{V}=f(x_i)\cdot\boldsymbol{W}_V^{\mathrm{T}} \tag{3}$$

其中,$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$ 分别代表 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的权重。在交叉注意力机制中,利用 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 得到缩放点积注意力之后的特征 $\hat{\boldsymbol{Z}}$:

$$\hat{\boldsymbol{Z}}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d'}}\right)\cdot\boldsymbol{V} \tag{4}$$

然后,将得到的特征 $\hat{\boldsymbol{Z}}$ 与其对应的类别语义向量进行加权融合进一步增强相关表示,从而得到解耦后的类别特征:

$$\boldsymbol{Z}^{\mathrm{SL}}=\alpha_1\cdot\hat{\boldsymbol{Z}}+(1-\alpha_1)\cdot g(x_i^{\mathrm{L}}) \tag{5}$$

式中,$\alpha_1=h(g(x_i^{\mathrm{L}}))$ 为可学习的参数。同理,利用交叉注意力机制捕捉视觉特征 $f(x_i)$ 和属性语义特征 $g(x_i^{\mathrm{A}})$ 的相关性,从而针对性地提取解耦后的属性特征,表示为:

$$\boldsymbol{Z}^{\mathrm{SA}}=\alpha_2\cdot\hat{\boldsymbol{Z}}+(1-\alpha_2)\cdot g(x_i^{\mathrm{A}}) \tag{6}$$

式中,$\alpha_2=h(g(x_i^{\mathrm{A}}))$ 也为可学习的参数。

通过上述过程,模型将重点关注视觉特征中与语义特征相关性较高的部分,通过交叉注意力机制提取视觉特征中与类别相关的特征和与属性相关的特征,并分别与原类别语义特征和原属性语义特征进行融合,从而进一步强化语义信息对视觉特征的指导作用,实现了视觉特征中类别与属性的解耦。

与支持集样本不同的是,查询集样本 x_q 不具备语义信息,因此采用自解耦的方法对查询集的视觉特征进行训练得到解耦后的类别特征。具体地,在利用交叉注意力机制时对于计算查询集样本的 $\boldsymbol{Q}'$、$\boldsymbol{K}'$和 $\boldsymbol{V}'$分别表示为:

$$\boldsymbol{Q}'=f(x_q)\cdot\boldsymbol{W}_{Q'}^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{K}'=f(x_q)\cdot\boldsymbol{W}_{K'}^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{V}'=f(x_q)\cdot\boldsymbol{W}_{V'}^{\mathrm{T}} \tag{9}$$

其中,$\boldsymbol{W}_{Q'}$、$\boldsymbol{W}_{K'}$ 和 $\boldsymbol{W}_{V'}$ 分别代表 $\boldsymbol{Q}'$、$\boldsymbol{K}'$ 和 $\boldsymbol{V}'$ 的权重。通过 $\boldsymbol{Q}'$、$\boldsymbol{K}'$和 $\boldsymbol{V}'$得到缩放的点积注意力之后的特征 $\hat{\boldsymbol{Z}}'$:

$$\hat{\boldsymbol{Z}}'=\mathrm{softmax}\left(\frac{\boldsymbol{Q}'\cdot\boldsymbol{K}'^{\mathrm{T}}}{\sqrt{d'}}\right)\cdot\boldsymbol{V}' \tag{10}$$

之后,通过残差连接与层归一化得到自解耦后的类别特征为 $\boldsymbol{Z}^{\mathrm{QL}}$:

$$\boldsymbol{Z}^{\mathrm{QL}}=LayerNorm(f(x_q)+\hat{\boldsymbol{Z}}') \tag{11}$$

1.4　特征重建

在此阶段,为保证支持集解耦后的类别特征与属性特征仍可准确表示原样本,采用特征重建的方式,通过将解耦后的特征进行融合得到重建后的特征,与原样本的全局特征进行约束。

具体而言,重建特征 $\boldsymbol{Z}_{\mathrm{rec}}^{\mathrm{S}}$ 表示为:

$$\boldsymbol{Z}_{\mathrm{rec}}^{\mathrm{S}}=\beta\cdot\boldsymbol{Z}^{\mathrm{SA}}+(1-\beta)\cdot\boldsymbol{Z}^{\mathrm{SL}} \tag{12}$$

式中,$\beta=h(\boldsymbol{Z}^{\mathrm{SA}})$ 同样是一个可学习的参数。

同时,原样本的全局特征由 $f(x_i)$ 池化得到:

$$\boldsymbol{Z}_{\mathrm{g}}^{\mathrm{S}}=p(f(x_i)) \tag{13}$$

式中,$p(\cdot)$ 为全局平均池化。之后对重建特征 $\boldsymbol{Z}_{\mathrm{rec}}^{\mathrm{S}}$和全局特征 $\boldsymbol{Z}_{\mathrm{g}}^{\mathrm{S}}$ 使用最小均方误差(mean squared error, MSE)进行约束:

$$\mathcal{L}_{\mathrm{rec}}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{Z}_{\mathrm{rec},i}^{\mathrm{S}}-\boldsymbol{Z}_{\mathrm{g},i}^{\mathrm{S}})^2 \tag{14}$$

因此模型学习特征信息时将继续朝着贴近样本本质的方向进行优化。

1.5　少样本训练

对于一个 N-way K-shot 的少样本分类任务,当 K 值不为 1 时,每个类别的支持集样本均采用原型网络[7]的方法,计算每个类别样本视觉特征、解耦后的类别特征和属性特征的向量均值作为对应的特征原型。以样本解耦的类别特征为例,对于类别 c,其解耦类别特征原型为:

$$\boldsymbol{P}_c^{\mathrm{L}}=\frac{1}{|X^{\mathrm{S}_c}|}\sum_{x\in X^{\mathrm{S}_c}}\boldsymbol{Z}^{\mathrm{SL}} \tag{15}$$

式中,X^{S_c} 表示支持集样本中属于类别 c 的数据,$|X^{\mathrm{S}_c}|$代表属于类别 c 的样本数量。

在给定一个距离度量函数 d 的条件下,根据得到的支持集解耦类别特征原型 $\boldsymbol{P}_c^{\mathrm{L}}$ 和查询集自解耦的类别特征 $\boldsymbol{Z}^{\mathrm{QL}}$,计算查询集样本 x_q 对于不同类别原型在特征空间中的类别分布:

$$p(y=c\,|\,x_q)=\frac{\exp(-d(\boldsymbol{Z}^{\mathrm{QL}},\boldsymbol{P}_c^{\mathrm{L}}))}{\sum_{c'}\exp(-d(\boldsymbol{Z}^{\mathrm{QL}},\boldsymbol{P}_{c'}^{\mathrm{L}}))} \tag{16}$$

这里,$p(y=c\,|\,x_q)$ 表示查询集样本 x_q 属于类别 c 的概率值,c'表示任一类别,$d(\cdot)$ 为欧氏距离的

相似度度量。

利用交叉熵损失作为其分类损失函数进行训练：

$$\mathcal{L}_{\text{cls}} = -\lg p(y=c\,|\,x_q) \tag{17}$$

合并所有损失函数得到最终的目标函数：

$$\mathcal{L}_{\text{obj}} = \arg\min \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{rec}} \tag{18}$$

式中，γ 为超参数，用来平衡最终的损失函数。

2 实验

2.1 实验设置

2.1.1 数据集

本文所提方法在 MIT-States[25] 和 C-GQA[26] 两个数据集上进行实验。MIT-States 数据集[25] 由 53 155 张现实世界图像组成，其中每个图像都包含它的形容词属性和名词类别，例如“黄色的苹果”。对该数据集按名词进行划分后，一共有 243 个名词类别，其中 160 个名词用作训练、40 个名词用于验证、43 个名词用来测试。此外，一共有 115 种形容词属性。同时，由属性和类别组成的词共 1 761 对，分别用在训练集、验证集和测试集的对数分别为 1 179、271 和 311。

C-GQA[26] 数据集源于 Stanford GQA 数据集[27]，具备更清晰的标签和更大的标签空间，挑战性更大。该数据集共有 27 062 张图像，按名词划分后得到训练集、验证集和测试集的图像数量分别为 20 786、3 511 和 2 765。相应地，其名词类别数量分别为 120、30、31。C-GQA 的训练集包含 153 种形容词属性。这里，由属性和类别组成的词共 931 对。数据集 MIT-States 和 C-GQA 的具体信息见表 1。

表 1 数据集划分

Tab. 1 Dataset splits

数据集	训练集				验证集				测试集			
	类别	属性	词对	图像	类别	属性	词对	图像	类别	属性	词对	图像
MIT-States	160	115	1 179	35 858	40	94	271	8 044	43	100	311	9 253
C-GQA	120	153	679	20 786	30	34	136	3 511	31	37	116	2 765

2.1.2 实验细节

在实验中，数据集 MIT-States 和 C-GQA 的图像均调整为 224 × 224 的尺寸，其类别和属性的语义向量通过 word2vec[28] 得到，维度为 600。并且，在两个数据集上训练并测试了 5-way 1-shot 和 5-way 5-shot 任务的准确率。首先，使用 Adam 优化器进行预训练，在训练集样本上训练 60 个批次。这里初始学习率设置为 0.001。之后，采用元训练的方式进行微调。在微调时，利用随机梯度下降（stochastic gradient descent, SGD）优化器[29] 训练 30 个批次，每个批次随机采样 1 000 个 episode；初始学习率设置为 0.000 1，并且每隔 10 个批次学习率减小至原来的 1/10；此时动量参数设置为 0.9，权重衰减率为 0.001。在训练和测试时，视觉特征的提取均采用 ResNet18[30] 作为主干网络，特征提取器的输出维度为 512；语义特征的提取则采用多层感知机（multilayer perceptron, MLP），其输出维度也为 512。对于超参数 γ，在 5-way 1-shot的任务中取 0.4，在 5-way 5-shot 任务中取 0.3。所有对比方法都使用相同的视觉和语义特征，并且调整参数达到其最优性能。

2.2 实验结果

与常见少样本分类工作的测试设置不同，将测试集分成两个部分：与支持集属性相同的查询集样本和与支持集属性不同的查询集样本。首先，在测试集中随机采样 2 000 个 episode，每个 episode 中的每个类别包含 2 个无标注的查询样本用来测试，即与支持集属性相同的测试样本（same support attribute, SSA）和与支持集属性不同的测试样本（different support attribute, DSA）各一个。然后，对 2 000 个 episode 上的 SSA 和 DSA 分别求平均准确率。最后，求所提方法对多元属性样本准确率，即求两种准确率的调和平均数（harmonic mean, HM）作为最终评价指标：

$$HM = 2 \times \frac{SSA \cdot DSA}{SSA + DSA} \tag{19}$$

本文选取了 7 种对比方法，包括 3 种不使用语义信息的方法（ProtoNet[7]、RelationNet[8] 和 MatchingNet[31]）以及 4 种使用语义信息的方法（AM3[11]、AGAM[12]、MAP-Net[23] 和 SGAP[32]）。表 2 和表 3 为本文方法在 2 个数据集上的实验结果。可以明显地看出，本文方法性能均较为显著地优于所有对比算法。另外，在 SSA 上的分类性能整体较 DSA 上的分类性能高，这表明多元属性问题确实会影响模型的分类性能。

表 2　在 MIT-States 数据集上的测试结果

Tab. 2　Test results on the MIT-States dataset

%

方法	5-way 1-shot			5-way 5-shot		
	SSA	DSA	HM	SSA	DSA	HM
ProtoNet[7]	54.10	45.85	49.63	69.57	55.01	61.03
RelationNet[8]	49.44	42.40	45.65	68.26	55.19	61.03
AM3[11]	61.53	53.89	57.46	69.94	60.24	64.73
AGAM[12]	55.76	49.05	52.19	66.65	43.72	60.10
MAP-Net[23]	62.44	51.80	56.62	71.70	57.96	64.10
MatchingNet[31]	52.22	44.80	48.23	66.87	54.04	59.77
SGAP[32]	32.75	27.95	30.16	61.94	49.99	55.33
MCD(本文)	**64.19**	**59.53**	**61.77**	**73.04**	**63.34**	**67.85**

表 3　在 C-GQA 数据集上的测试结果

Tab. 3　Test results on the C-GQA dataset

%

方法	5-way 1-shot			5-way 5-shot		
	SSA	DSA	HM	SSA	DSA	HM
ProtoNet[7]	39.24	34.69	36.82	54.33	46.85	50.31
RelationNet[8]	43.54	38.36	40.79	59.85	51.04	55.10
AM3[11]	50.51	44.47	47.30	56.82	48.84	52.53
AGAM[12]	46.09	41.41	43.62	58.67	50.74	54.42
MAP-Net[23]	50.05	43.85	46.75	62.77	52.32	57.07
MatchingNet[31]	44.60	39.48	41.88	58.24	49.07	53.26
SGAP[32]	36.99	29.92	33.08	53.98	44.81	48.97
MCD(本文)	**52.50**	**47.91**	**50.10**	**64.78**	**56.34**	**60.27**

具体地,在 MIT-States 数据集上,与结果第二高的比较方法 AM3 或 MAP-Net 相比,对于 5-way 1-shot少样本分类任务,本文方法在 SSA 的结果上提升了 1.75%,在 DSA 上的测试结果提高了 5.64%,并且调和平均数 HM 也有了 4.31% 的提升效果;在 5-way 5-shot 任务中,本文方法在 SSA 上比方法 MAP-Net 提高了 1.34%,在 DSA 上比方法 AM3 提高了 3.10%,在 HM 上也有 3.12% 的提升效果。在 C-GQA 数据集上,针对 5-way 1-shot的分类结果,本文方法比结果第二高的方法 AM3 均有明显的提高,在 SSA 上提高 1.99%,在 DSA 上有 3.44% 的提升,并且在 HM 上也有 2.80% 的提升效果;而在 5-way 5-shot 任务中,本文方法仍比结果第二高的方法 MAP-Net 有较高的提升,其中在 SSA 上提升了 2.01%,在 DSA 上提高了 4.02%,在 HM 上也有 3.20% 的明显提升。

通过实验结果还可以发现,本文方法在 DSA 上较对比方法的提升效果明显比在 SSA 上显著,这说明对比方法难以很好地对具有未见属性的同类样本进行准确识别,本文方法在一定程度上有效地缓解了这一问题。同时,与5-way 5-shot提升结果相比,5-way 1-shot 的分类任务中,DSA 的提升效果在两个数据集上都更为明显。主要的原因在于当支持集样本极少时,模型更难学习样本的类别表示信息,并且当属性差异较大时,更容易因属性不同混淆本质的类别表征,加剧识别类别错误的情况。本文方法能够有效地解耦出无关的属性信息,保留样本本质的类别特征,提高样本极少情况下的分类性能。

2.3 消融实验

如前所述,本文方法通过交叉解耦和特征重建两个关键模块提升多元属性少样本分类性能。接下来,为了探究每个模块对实验结果的影响,在MIT-States和C-GQA数据集上进行消融实验研究,结果如表4和表5所示。采用原型网络作为基准即第一行结果,第二行为仅增加支持集类别语义特征(support category features, SCF),第三行的结果为在此基础上增加支持集属性语义特征(support attribute features, SAF)形成交叉解耦模块,第四行为继续增加特征重建模块形成本文最终方法。从表中结果可以发现交叉解耦与特征重建两个模块分别对少样本的分类性能起到促进作用。

表4 在MIT-States数据集上的消融实验结果

Tab.4 Ablation study results on the MIT-States dataset

%

方法	5-way 1-shot			5-way 5-shot		
	SSA	DSA	HM	SSA	DSA	HM
ProtoNet[7]	54.10	45.85	49.63	69.57	55.01	61.03
ProtoNet + SCF	59.88	55.81	59.27	69.54	60.96	65.57
ProtoNet + SCF + SAF	63.16	58.20	60.58	72.14	61.85	66.60
MCD	**64.19**	**59.53**	**61.77**	**73.04**	**63.34**	**67.85**

表5 在C-GQA数据集上的消融实验结果

Tab.5 Ablation study results on the C-GQA dataset

%

方法	5-way 1-shot			5-way 5-shot		
	SSA	DSA	HM	SSA	DSA	HM
ProtoNet[7]	39.24	34.69	36.82	54.33	46.85	50.31
ProtoNet + SCF	49.64	46.01	47.76	62.24	54.09	57.88
ProtoNet + SCF + SAF	51.11	46.98	48.96	63.08	55.05	58.79
MCD	**52.50**	**47.91**	**50.10**	**64.78**	**56.34**	**60.27**

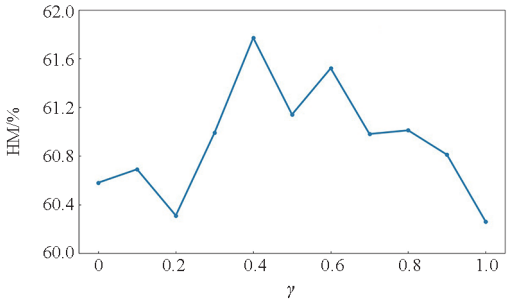
具体地,当仅引入类别语义信息时,与基准原型网络方法相比,测试结果几乎都有显著的提升,这表明增加类别语义特征能够弥补模型学习视觉特征的不足,对模型学习准确分类起到了积极的指导作用。

当加入属性语义信息使用交叉融合模块时,可以发现交叉融合的方法在SSA、DSA和HM上的分类性能大部分都比仅加入类别语义特征SCF时有较明显的提升,其提升效果能达到0.84%~3.28%。特别地,当测试样本的属性特征与支持集样本相同时,SSA的结果反映出交叉解耦模型可以更好学习出解耦后的类别特征和属性特征,分离出正确的类别特征用以分类,有效地提高少样本任务分类的准确率,其性能能够达到0.84%~3.28%的提升。同时,即使在测试样本的属性特征与支持集样本不同时,模型也能学习出未见属性特征并解耦出类别特征。在此基础上当继续引入特征重建模块时,SSA、DSA和HM的性能均存在进一步的提升。通过结果可以发现,在解耦过程中,可能会出现解耦出来的特征偏离原样本真实特征的情况。因此,特征重建模块将解耦后的类别特征和属性特征重新融合,并利用原样本视觉特征进行约束,使模型学习特征信息时沿着样本类别内在特征的方向进行优化。

2.4 进一步分析

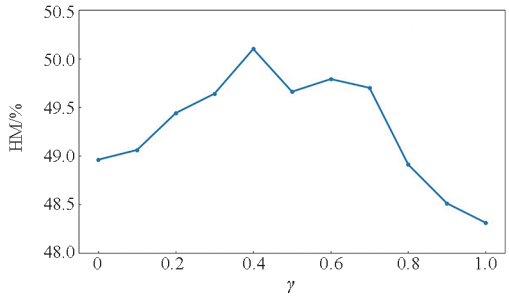
2.4.1 参数实验

在此部分,研究超参数γ对模型性能的影响。不同γ值在MIT-States数据集和C-GQA数据集的结果如图4所示,其中图4(a)和图4(c)表示MIT-States数据集的结果,图4(b)和图4(d)表示C-GQA数据集的结果。蓝色和黄色折线分别代表5-way 1-shot和5-way 5-shot少样本分类任务,采用调和平均数HM作为判别指标。



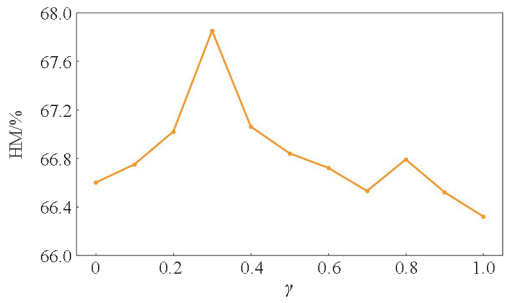
(a) MIT-States 数据集在 5-way 1-shot 的结果

(a) Results of 5-way 1-shot on the MIT-States dataset



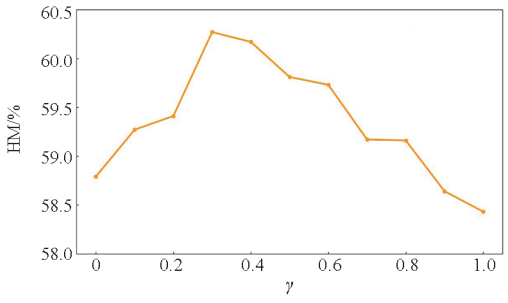
(b) C-GQA 数据集在 5-way 1-shot 的结果

(b) Results of 5-way 1-shot on the C-GQA dataset



(c) MIT-States 数据集在 5-way 5-shot 的结果

(c) Results of 5-way 5-shot on the MIT-States dataset



(d) C-GQA 数据集在 5-way 5-shot 的结果

(d) Results of 5-way 5-shot on the C-GQA dataset

图 4　超参数 γ 对性能的影响

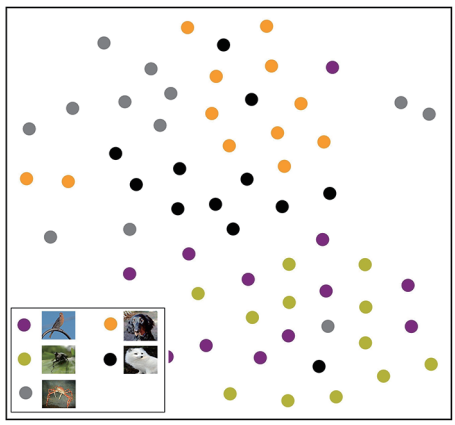
Fig. 4　Impact of hyper-parameter γ on performance

通过实验发现,两种数据集上的超参数 γ 均是先升后降。在 5-way 1-shot 分类任务中,当 γ = 0.4 时,网络性能达到最优,HM 的结果最佳;而在 5-way 5-shot 少样本任务中,当 γ = 0.3 时,模型能达到最佳性能,此时 HM 的值为最高。这表明特征重建模块能够在一定程度上有效地提升分类性能,但是当参数 γ 的值过大时,分类性能反而会下降,其原因可能在于重建约束较强时会影响属性解耦的效果,加重了属性对类别特征的影响。

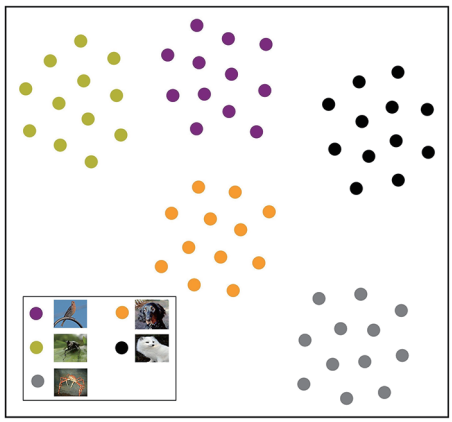
2.4.2　可视化实验

为了验证本文方法的效果,采用t-SNE[33]方法在 MIT-States 数据集的嵌入空间进行可视化实验,结果如图 5 所示。随机可视化一个测试任务,这个任务包含 5 个类别共 60 个样本,每个类别的样本数为 12。不同颜色的圆点代表不同的类别。这里,选用原型网络(ProtoNet)作为基准方法进行比较,结果如图 5(a)所示;为了验证本文方法的可行性,图 5(b)为仅引入类别语义信息的方法结果,图 5(c)为本文方法结果。

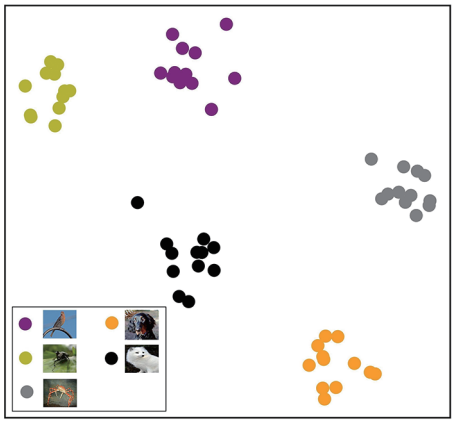
从 t-SNE 的可视化结果可以看出,增加类别语义信息,能够有效地提高类别辨识的能力,使测试样本按照预测类别整齐规则分布。本文方法进一步增强了同类别样本的分布密度,使类内样本分布更加紧凑,同时提高类别间的距离,可降低多元属性问题带来的识别误差。



(a) ProtoNet



(b) ProtoNet + SCF



(c) MCD

图 5 MIT-States 数据集的 t-SNE 可视化结果

Fig. 5 t-SNE visualization on MIT-States dataset

3 结论

本文针对少样本学习中的多元属性问题,提出一种多模态交叉解耦的方法,可有效解耦出样本类别特征和类别属性语义特征,并能利用特征重建保留样本内在的类别特征信息,极大地缓解了多元属性问题对少样本分类造成的识别错误。所提方法也提出一种特征重建方法,将解耦后的类别特征和属性特征重新融合,通过原样本的视觉特征进行约束,使模型学习特征信息时沿着贴近样本本质类别的方向进行优化。在两个类内属性差异较大的标准数据集上的大量实验验证了所提方法的有效性和先进性,结果显示本文方法能够有效提升少样本分类任务的准确性,即使在属性未知的待测样本中也具有良好的分类性能。

参考文献(References)

[1] 刘茂华, 韩梓威, 陈一鸣, 等. 机载激光雷达数据的三维深度学习树种分类[J]. 国防科技大学学报, 2022, 44(2): 123 – 130.
LIU M H, HAN Z W, CHEN Y M, et al. Tree species classification of airborne LiDAR data based on 3D deep learning[J]. Journal of National University of Defense Technology, 2022, 44(2): 123 – 130. (in Chinese)

[2] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// Proceedings of the 34th International Conference on Machine Learning, 2017: 1126 – 1135.

[3] 王春哲, 安军社, 姜秀杰, 等. 融合神经网络与超像素的候选区域优化算法[J]. 国防科技大学学报, 2021, 43(4): 145 – 155.
WANG C Z, AN J S, JIANG X J, et al. Region proposals optimization algorithm combining neural networks and superpixels[J]. Journal of National University of Defense Technology, 2021, 43(4): 145 – 155. (in Chinese)

[4] KANG B Y, LIU Z A, WANG X, et al. Few-shot object detection via feature reweighting[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 8419 – 8428.

[5] 张艳, 相旭, 唐俊, 等. 跨模态行人重识别的对称网络算法[J]. 国防科技大学学报, 2022, 44(1): 122 – 128.
ZHANG Y, XIANG X, TANG J, et al. Cross-modality person re-identification algorithm using symmetric network[J]. Journal of National University of Defense Technology, 2022, 44(1): 122 – 128. (in Chinese)

[6] WU L, WANG Y, YIN H Z, et al. Few-shot deep adversarial learning for video-based person re-identification[J]. IEEE Transactions on Image Processing, 2020, 29: 1233 – 1245.

[7] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 4080 – 4090.

[8] SUNG F, YANG Y X, ZHANG L, et al. Learning to compare: relation network for few-shot learning[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2018: 1199 – 1208.

[9] SUN Q R, LIU Y Y, CHUA T S, et al. Meta-transfer learning for few-shot learning[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 403 – 412.

[10] YU Y L, ZHANG D Y, WANG S D, et al. Local spatial alignment network for few-shot learning[J]. Neurocomputing, 2022, 497: 182 – 190.

[11] XING C, ROSTAMZADEH N, ORESHKIN B N, et al. Adaptive cross-modal few-shot learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 4847 – 4857.

[12] HUANG S T, ZHANG M, KANG Y C, et al. Attributes-guided and pure-visual attention alignment for few-shot recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(9): 7840 – 7847.

[13] RUSU A A, RAO D, SYGNOWSKI J, et al. Meta-learning with latent embedding optimization[C]//Proceedings of the International Conference on Learning Representations, 2018.

[14] CHEN Z T, FU Y W, WANG Y X, et al. Image deformation meta-networks for one-shot learning[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 8672 – 8681.

[15] LI K, ZHANG Y L, LI K P, et al. Adversarial feature hallucination networks for few-shot learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 13467 – 13476.

[16] LI W B, WANG L, XU J L, et al. Revisiting local descriptor based image-to-class measure for few-shot learning[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 7253 – 7260.

[17] CHEN B H, DENG W H. Hybrid-attention based decoupled metric learning for zero-shot image retrieval[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 2745 – 2754.

[18] CHEN H, DING G G, LIU X D, et al. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), 2020: 12655 – 12663.
[19] WANG H R, ZHANG Y, JI Z, et al. Consensus-aware visual-semantic embedding for image-text matching[C]// Proceedings of European Conference on Computer Vision, 2020: 18 – 34.
[20] DIAO H W, ZHANG Y, MA L, et al. Similarity reasoning and filtration for image-text matching[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1218 – 1226.
[21] JI Z, SUN Y X, YU Y L, et al. Attribute-guided network for cross-modal zero-shot hashing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(1): 321 – 330.
[22] MANCINI M, NAEEM M F, XIAN Y Q, et al. Open world compositional zero-shot learning[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 5218 – 5226.
[23] JI Z, HOU Z S, LIU X Y, et al. Information symmetry matters: a modal-alternating propagation network for few-shot learning[J]. IEEE Transactions on Image Processing, 2022, 31: 1520 – 1531.
[24] ZHU Y H, MIN W Q, JIANG S Q. Attribute-guided feature learning for few-shot image recognition[J]. IEEE Transactions on Multimedia, 2021, 23: 1200 – 1209.
[25] ISOLA P, LIM J J, ADELSON E H. Discovering states and transformations in image collections[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1383 – 1391.
[26] NAEEM M F, XIAN Y Q, TOMBARI F, et al. Learning graph embeddings for compositional zero-shot learning[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 953 – 962.
[27] HUDSON D A, MANNING C D. GQA: a new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 6693 – 6702.
[28] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 3111 – 3119.
[29] ROBBINS H, MONRO S. A stochastic approximation method[J]. The Annals of Mathematical Statistics, 1951, 22(3): 400 – 407.
[30] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770 – 778.
[31] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 3637 – 3645.
[32] JI Z, LIU X Y, PANG Y W, et al. Few-shot human-object interaction recognition with semantic-guided attentive prototypes network[J]. IEEE Transactions on Image Processing, 2021, 30: 1648 – 1661.
[33] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579 – 2605.

(编辑: 梁慧, 杨琴)