

基于对比学习文本感知的小样本 遥感图像分类

李文辉, 杨 晨

(吉林大学 计算机科学与技术学院, 长春 130012)

摘 要: 针对现有方法主要利用遥感图像单一模态解决同种类别相似度低的问题, 本文提出了一种基于多模态学习的遥感图像分类方法。首先, 修正图像的空间特征并利用对比学习进行图像编码器的预训练进而生成图像特征, 利用文本编码器生成文本特征。其次, 引入特征解码器获取文本感知的视觉特征, 在特征融合阶段提出了一个新的注意力机制方法。再次, 设计了一个新的图像编码器用于提高分类精度。最后, 通过计算支持集和查询集之间的相似性进一步进行类别预测。在 NWPU-RESISC45、AID 和 UC Merced 数据集上进行实验, 其 5-way 5-shot 准确度分别达到 86.46%、85.89% 和 80.32%, 优于现有的小样本遥感图像分类方法。

关键词: 计算机应用技术; 对比学习; 文本感知; 小样本学习; 遥感图像分类

中图分类号: TP751.1 **文献标志码:** A **文章编号:** 1671-5497(2025)07-2393-09

DOI: 10.13229/j.cnki.jdxbgxb.20231128

Few-shot remote sensing image classification based on contrastive learning text perception

LI Wen-hui, YANG Chen

(College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Abstract: Aiming at the problem that existing methods mainly use a single modality of remote sensing images to solve the problem of low similarity of the same class, a remote sensing image classification method based on multimodal learning is proposed. Firstly, The spatial features of the image are corrected. The image encoder is pre-trained using contrastive learning to generate image features and text features are generated using text encoder. Secondly, a feature decoder is introduced to acquire text perception visual features, and a new attention mechanism approach is proposed in the feature fusion stage. Thirdly, a new image encoder is designed to improve the classification accuracy. Finally, the similarity between the support set and the query set is computed for further class prediction. Experiments are conducted on the NWPU-RESISC45, AID and UC Merced datasets. The 5-way 5-shot accuracies of 86.46%, 85.89%, and 80.32% respectively outperform existing methods in few-shot remote sensing image classification.

Key words: computer application technology; contrastive learning; text perception; few-shot learning; remote sensing image classification

收稿日期: 2023-10-20.

基金项目: 吉林省科技发展计划项目(20230201082GX).

作者简介: 李文辉(1961-), 男, 教授, 博士. 研究方向: 计算机图形学, 计算机视觉. E-mail: liwh@jlu.edu.cn

0 引言

遥感图像分类是对输入的遥感图像采用一定的方法进行分类的过程。这些技术已广泛应用于城市土地利用类型的精确分类、资源勘探和自然灾害的检测^[1,2]。目前,随着卫星遥感技术的不断进步,高分辨率场景的图像数量不断增加。由于标注数据的人力物力成本很高,传统的基于深度卷积网络^[3]的遥感图像分类方法受限于大量的人工标注数据进行训练,所以减少大量的人工标注数据进行训练的研究受到关注^[2,4]。

小样本学习是指利用先验知识快速识别新类别的能力,即使每个新类别只有少数标记图像可用^[5]。现有的小样本遥感图像分类方法采用两个阶段进行训练和测试。第一阶段将网络预训练为嵌入网络,第二阶段用度量学习进行分类任务。Snell等^[6]提出了原型网络,其思想是在神经网络学习的表示空间中通过其示例的平均值表示每个类。Sung等^[7]提出了关系网络,能够通过计算查询图像与每个新类别的少量示例之间的关系得分对新类别的图像进行分类,而无须进一步更新网络。Li等^[1]提出了自适应匹配网络的判别性学习,将支撑集合和查询集合的判别性特征沿深度拼接,利用一个匹配器适应性地选择语义相关的样本对,并计算相似度得分。Ye等^[8]提出了一种带有Transformer的小样本嵌入适应方法,通过一种集合到集合的函数将实例嵌入适应目标分类任务,从而产生任务特定且具有区分能力的嵌入。Cheng等^[9]提出了原型自校准和相互校准的连体原型网络,自校准利用支撑集合的标签提供的监督信息校准支撑特征的原型,互校准将查询集合里的样本的置信度得分作为另一种形式的原型,用同样的方式预测支撑集样本。

近年来,随着自监督学习的不断发展,自监督学习技术已经被应用于遥感图像分类。这些方法使模型能够学习数据中的底层结构和关系,减少对标记数据的依赖,提高分类性能。Li等^[2]提出了基于自监督对比学习的度量学习框架,将自监督对比学习融入小样本中,能够让特征提取网络在较少数据的情况下依然学习到具有代表性的图像特征。虽然对小样本遥感图像分类任务的大部分研究都集中在可区分的特征上,但研究者主要是利用图像单一模态。随着多模态学习的发展,研究显示了多模态模型在预训练中的有效性,尤

其是视觉文本模型^[10]。因此,如果能够利用多模态交互提示分类器,特征提取的效果将得到提升。

因此,本文提出一种基于对比学习文本感知的度量学习框架 CLTP-ML(Contrastive learning text perception-metric learning),通过优化图像特征和文本特征的相似度关系,并得到特征的高级融合表达,进而提升遥感图像分类任务的准确度。

1 本文方法

为了模型能够有效捕捉到区分类的特征,本文提出基于对比学习文本感知的度量学习框架 CLTP-ML。首先,简要介绍预训练阶段对比学习的整个过程;然后,介绍整个框架中的特征编码器以及改进的图像编码器具体结构,再详细介绍特征解码器生成文本感知的视觉特征以及新提出的注意力机制结构;最后,介绍度量学习阶段的过程。

1.1 预训练阶段

在预训练阶段,为了训练出能够学习增强特征表示的图像编码器,本文采用基于对比学习的方法从训练集 D_{tr} 中抽取 N 对图像和标签,并对每个样本进行扩充,总共得到 $2N$ 个增强样本。首先给出两个样本对之间损失函数计算公式:

$$L_{ij} = -\log \frac{\exp(\cos(f_i, f_j)/\tau)}{\sum_{k=1}^{2N} \exp(\cos(f_i, f_k)/\tau)} \quad (1)$$

式中: $i \neq k$ 时上述等式成立, τ 为温度系数。则对全局损失函数^[11]可以定义为:

$$L_g = \sum_{i=1}^{2N} \frac{1}{2N_{y_i} - 1} \sum_{j=1}^{2N} L_{ij} \quad (2)$$

式中: $i \neq j$ 且 $y_i = y_j$ 时等式成立。

虽然全局损失都能够让图像编码器产生高级特征表达,但是全局损失也会造成对局部区域的判别丢失情况。针对这种情况,在预训练阶段加入一种基于自注意力机制的对比学习方法。图1为整个预训练对比学习过程。假定给出两张遥感图像,通过图像编码器得到对应的空间特征向量。为了计算二者之间的局部损失,采用自注意力机制^[12]对它们进行空间对齐。设 k, q, v 分别为 d 维度的键向量、查询向量和值向量,给定一对空间特征向量 f_a 和 f_b 作为输入,要计算 a 相对于 b 的对齐值 $v_{a|b}$ 需要找到 a 和 b 两个特征中对相似度影响大的那些空间位置,并体现在相似度矩阵 $\mathcal{D}_{ab}$ 中:

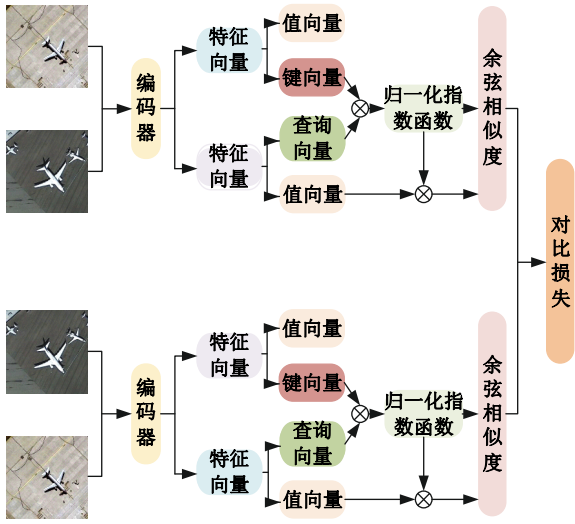


图 1 预训练对比学习阶段

Fig. 1 Pre-training contrastive learning stage

$$\partial_{ab} = \text{Soft max} \left(\frac{q_b k_a^T}{\sqrt{d}} \right) \quad (3)$$

式中:Soft max 为图中的归一化指数函数。接下来,再使 a 突出这些有更大影响力的空间位置处的特征,完成特征对齐,则给出对齐值 $v_{a|b}$ 的计算公式:

$$v_{a|b} = \partial_{ab} v_a \quad (4)$$

同理还会计算一次 b 相对于 a 的对齐值 $v_{b|a}$ 。因此,计算一对空间特征之间的相似性的表达式如下:

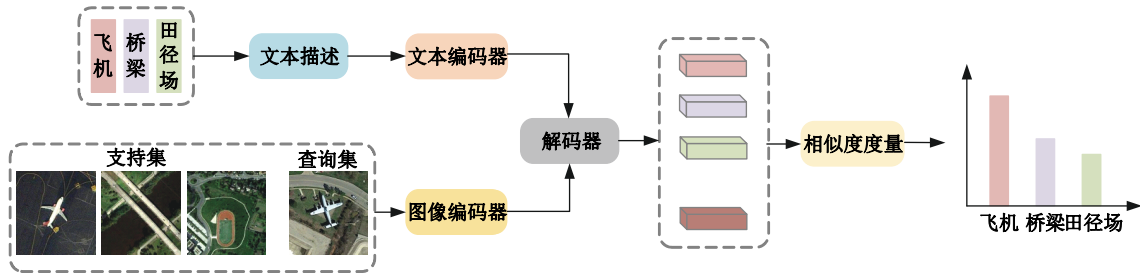


图 2 CLTP-ML 框架结构图

Fig. 2 Structure of CLTP-ML framework

CLTP-ML 框架中的图像编码器 MRNet 共享预训练对比学习阶段的参数。具体网络结构是基于传统的 ResNet12 模型^[15]的基础上引入了 Mixer^[16]架构。如图 3 所示,首先将遥感图像输入 ResNet12 模型中,以第一个残差块为例,通过残差块后对输出张量进行维度转换,然后分别通过层归一化、线性层、GeLU 激活函数和线性层。将其结果与输入张量相加后通过一次层归一化,接下来与通过线性层、GeLU 激活函数、线性层的结

$$\text{sim}(f_a, f_b) = \frac{1}{HW} \sum_{r=1}^{HW} [v_a^T v_{b|a} + v_b^T v_{a|b}] \quad (5)$$

每个空间位置都按上述方式计算一次相似度,再将结果相加取平均。则可以定义局部损失函数为:

$$L_l = \sum_{a=1}^{2N} \frac{1}{2N_{y_a} - 1} \sum_{b=1}^{2N} L_{ab} \quad (6)$$

L_{ab} 的表达式为:

$$L_{ab} = -\log \frac{\exp(\text{sim}(f_a, f_b)/\tau)}{\sum_{k=1}^{2N} \exp(\text{sim}(f_a, f_k)/\tau)} \quad (7)$$

针对预训练阶段的目标主要考虑的是全局损失函数和局部损失函数的加权系数和,分别用 μ 和 λ 控制每一项的贡献,后续实验中将讨论具体系数的选择。预训练阶段总损失函数定义为:

$$L = \mu L_g + \lambda L_l \quad (8)$$

1.2 特征编码器

图 2 为 CLTP-ML 框架的结构图。在小样本学习的数据准备中,通常将数据集分为支持集 S 和查询集 Q 。CLTP-ML 框架中的文本编码器采用的与 CLIP^[13]训练的文本编码器一致,具体的文本信息映射描述与文献[14]一致。首先,将每张图像类别标签生成对应的文本描述。其次,将文本描述输入文本编码器中。文本编码器用于将自然语言文本转换为抽象的特征表示,特征向量能够有效捕捉到文本的语义和上下文信息。

果相加得到第一个残差块的结果。上述操作一共进行 4 轮,分别通过 ResNet12 中的 4 个残差块。

1.3 特征解码器

由于遥感图像存在类内相似性低的问题,使从有限样本中提取可区分特征变得更具挑战性。为了解决这个问题,通过将图像的视觉特征和文本的语义特征相结合,可以增加样本的多样性和信息丰富性,从而提高对不同类别的区分能力。因此,CLTP-ML 框架采用一种基于多头注意力

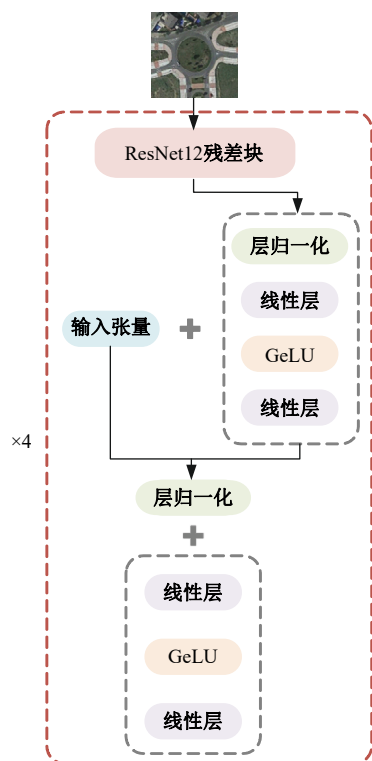


图3 图像编码器结构图

Fig. 3 Structure of image encoder

机制^[12]的特征解码器生成文本感知的视觉特征。图4为特征解码器结构图。首先将生成的文本特征和图像特征分别进行层归一化和线性层的操作得到 T_1 和 I_1 , 然后将特征投影到同一空间中利用多头注意力机制对视觉特征进行校正, 将其重新映射到视觉空间, 具体融合过程如图5所示。首先将 I_1 输入层归一化和多头自注意力机制中, 利用文本特征 T_1 作为键和值获得文本感知特征 I , 然后将其重新映射到视觉空间。在进行文本特征图像特征融合和对齐时, 多头自注意力将输入的序列数据分成多个头, 对每个头进行独立的查询、键、值线性变换, 可表示为:

$$\text{head}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (9)$$

式中: QW_i^Q, KW_i^K, VW_i^V 分别为第 i 个头的查询、键、值变换矩阵, 然后对每个头进行自注意力计算:

$$\text{attention}(Q, K, V) = \text{Soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

式中: d_k 为键向量的维度。最后将所有头的输出拼接在一起, 并进行输出线性变换, 具体如下式所示:

$$\text{multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^O \quad (11)$$

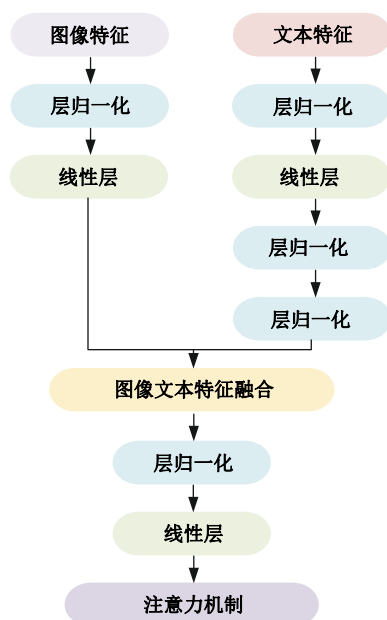


图4 特征解码器结构图

Fig. 4 Structure of feature decoder

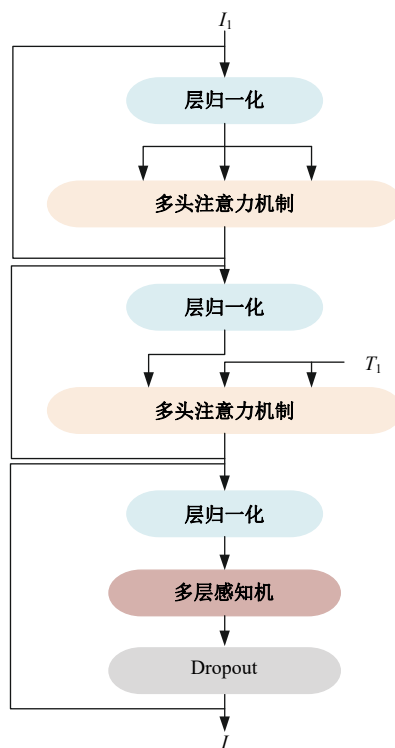


图5 图像文本特征融合

Fig. 5 Fusion of image features and text features

式中: n 为头的数量; W^O 为输出变换矩阵。

最后利用新设计的注意力机制计算空间和通道注意力图, 增强感兴趣的特征。图6为新设计的注意力机制 ECSAM 结构图。首先在 H 和 W 上对输入进来的特征层进行最大池化和平均池化, 然后进行一个 1D 卷积再取一次 Sigmoid 将值固

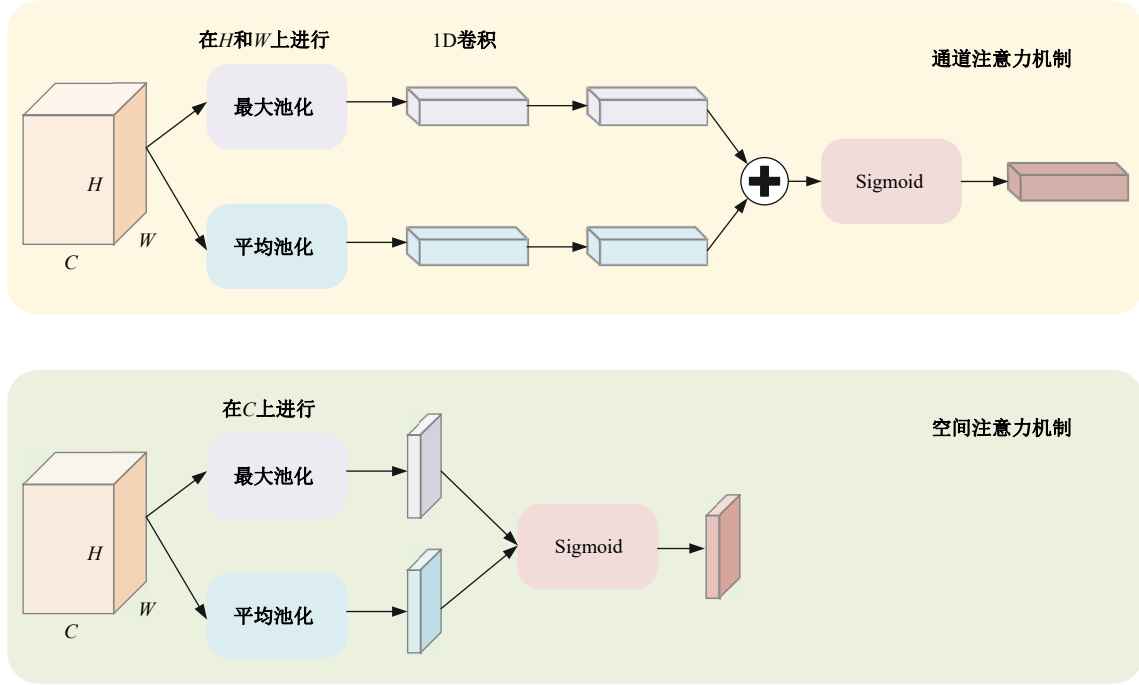


图 6 注意力机制结构图

Fig. 6 Structure of attention mechanism

定在 0~1,最后将这个权值乘上原输入特征层。进行通道注意力机制后,进行空间注意力机制。首先对输入进来的特征层,在每一个特征点的通道上取最大值和平均值。之后将这两个结果进行一个堆叠,利用一次通道数为 1 的卷积调整通道数,然后取一个 Sigmoid,将这个权值乘上原输入特征层。

1.4 度量学习阶段

假设每个类别的训练样本向量的平均值设为 m ,那么 k 类训练样本的质心表示为:

$$m_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S} f(x_i) \quad (12)$$

式中: S_k 为标记 k 类的支持集样本。在测试阶段,对于每个测试样本,计算它与所有类别的原型之间的相似度,并将测试样本分配给与其相似性最高的类别。本文选择的是余弦相似度作为距离函数 $d(\cdot)$,为了在所有类中产生一个分布,从查询集 Q 中选取一个查询样本 q ,则其分布表达式为:

$$p(y = k|q) = \frac{\exp(-d(f(q), m_k))}{\sum_{k_1=1}^c \exp(-d(f(q), m_{k_1}))} \quad (13)$$

因此,度量学习阶段的总损失函数为:

$$L_{\text{Test}} = d(f(q), m_k) + \log \sum_{k_1=1}^c d(f(q), m_{k_1}) \quad (14)$$

2 实验结果与分析

2.1 数据集

为验证 CLTP-ML 在遥感图像分类任务中的有效性,先后在 3 个公共遥感数据集上进行实验,主要包括 NWPU-RESISC45 数据集^[17]、AID 数据集^[18]和 UC Merced 数据集^[19]。NWPU-RESISC45 数据集是西北工业大学公布的用于遥感图像场景分类的大规模公开数据集,包含 45 类场景。每个类别各包含 700 张遥感图像,整个数据集一共 31 500 张图像,每张图像的像素大小为 256×256 像素,具有规模大、图像中所含信息丰富等特点。AID 数据集是一个大型航空图像数据集,从 Google Earth 图像中收集样本图像,包含 30 个场景类,每个类有 220~420 张图像。整个数据集一共有 10 000 张图像,每张图像的像素大小为 600×600 像素。UC Merced 数据集是从美国地质勘探局国家地图城市地区图像数据库中收集的,包含 21 个不同的场景类,每个类包含 100 张图像。整个数据集一共有 2 100 张图像,每张图像的像素大小为 256×256 像素。

2.2 实验环境和参数设置

以 PyCharm 为实验平台,使用 PyTorch 开发框架实现模型,该模型部署在带有 CUDA 版本 11.4 的 NVIDIA GeForce RTX 3090 GPU 上。

模型使用 Adam 优化器进行优化,学习率为 0.000 1,学习率调整乘数为 0.5,学习率衰减区间为 10,温度系数为 128。

2.3 图像预处理

NWPU 数据集随机选取 25 个类用于训练数据集,10 个类用于验证数据集以及 10 个类用于测试数据集。AID 数据集随机选取 16 个类用于训练数据集,7 个类用于验证数据集以及 7 个类用于测试数据集。UCM 数据集随机选取 10 个类用于训练数据集,5 个类用于验证数据集以及 6 个类用于测试数据集。为了进行有效的实验比较,划分方式遵循 SCL-MLNet^[2]的划分。为了公平地比较,实验中采用的图像编码器为新设计的图像编码器。采用的数据增强方式为随机水平翻转,随机裁剪和随机灰度变换。实验时随机生成 100 集用于训练,500 集用于验证,2 000 集用于测试。在训练和测试阶段,用 15 个查询样本评估模型在 5-way 5-shot 情况下的性能。

2.4 总体实验结果及比较

本文将 CLTP-ML 框架与一些经典和前沿的小样本遥感图像分类方法进行对比,包括原型网络(ProtoNet)^[6]、关系网络(RelationNet)^[7]、自适应匹配网络(DLA-MatchNet)^[1]、自监督对比学习度量学习框架(SCL-MLNet)^[2]、带有 Transformer 的小样本嵌入适应方法(FEAT)^[8]、原型自校准和相互校准的连体原型网络(SPNet)^[9]及文本感知框架(TeAw)^[10]。表 1 显示了现有的小样本遥感图像分类模型与本文框架的分类准确度对比。在 NWPU 数据集中,CLTP-ML 框架在 5-way 5-shot 情况下的分类准确率达到 86.46%,与 TeAw 方法相比高出 0.89%。在 AID 数据集中,CLTP-ML 框架在 5-way 5-shot 情况下的准确率达到了 85.89%,比排名第二的 TeAw 方法相比高出 1.27%。在 UCM 数据集中,5-way 5-shot 情况下的准确率可以达到 80.32%,相较于 TeAw 方法相比提升 2.82%。通过对比可以看出,本文提出的框架能够在小样本遥感分类任务中取得较好的效果。

2.5 参数实验结果及比较

本文对各个模型参数进行了实验,实验结果如表 2 所示。从实验结果中可以看出,参数量最小的模型是 ProtoNet,仅为 11.84 M,尽管参数量较小,但是 ProtoNet 在遥感图像分类任务上的准

表 1 各模型在 3 个数据集的分类准确度比较

Table 1 Comparison of classification accuracies of models on three datasets

模型	5-way 5-shot		
	NWPU	AID	UCM
ProtoNet ^[6]	66.64±0.51	70.33±0.48	67.84±0.15
RelationNet ^[7]	81.43±0.73	74.46±0.87	62.24±0.54
DLA-MatchNet ^[1]	82.62±0.54	76.57±0.74	63.86±0.37
SCL-MLNet ^[2]	82.77±0.29	79.83±0.66	67.39±0.74
FEAT ^[8]	83.84±0.37	81.39±0.71	75.87±0.52
SPNet ^[9]	80.55±0.46	78.28±0.58	73.52±0.53
TeAw ^[10]	85.57±0.25	84.62±0.25	77.50±0.27
CLTP-ML	86.46±0.43	85.89±0.24	80.32±0.06

率是最低的。FEAT 模型的参数量最大,达到 37.52 M。尽管 FEAT 方法的准确度有一定的提升,但是该方法的参数量大约达到了 ProtoNet 方法的 3 倍。在计算复杂度方面,RelationNet 的数值是最低的,FEAT 方法的数值是最高的。然而 CLTP-ML 方法结合参数量小、计算复杂度小以及准确度高几方面因素能够取得相对优越的性能。

表 2 各模型参数实验结果

Table 2 Experimental results for each model parameter

模型	参数量/M	FLOPs/G
ProtoNet ^[6]	11.84	1.82
RelationNet ^[7]	22.43	2.44
DLA-MatchNet ^[1]	15.06	1.95
SCL-MLNet ^[2]	14.24	2.03
FEAT ^[8]	37.52	3.98
SPNet ^[9]	12.96	2.26
TeAw ^[10]	14.59	2.17
CLTP-ML	14.35	2.06

2.6 预训练损失函数加权系数结果比较

为了使预训练阶段的训练效果达到最优,对预训练阶段总损失函数的加权系数进行讨论。如图 7 所示为不同加权系数在 3 个数据集上 5-way 5-shot 的分类准确度,横轴表示的是全局损失与局部损失的比值,即 μ/λ ,纵轴表示的是分类准确度。本文选择了 12 组不同情况的取值,前 9 组 μ 和 λ 的和等于 1,后 3 组 μ 确保等于 1,调整 λ 的值。从图中结果可以看出,对于前 9 组数据,随着 μ 的值不断减少 0.1, λ 的值不断增加 0.1,在 3 个数据集上 5-way 5-shot 的分类准确度不断下降。对于后 3 组数据,可以从图中看出,在 3 个数据集上的

折线走势都是先上升后降低,所有折线在 $\mu=1$ 、 $\lambda=0.5$ 处达到峰值且高于前9组中的最大数据。从结果中可以分析出,在全局损失和局部损失中全局损失占的比例更大,适当引入局部损失可以有效地提升模型的分类准确度。因此,选择1作为全局损失的系数,选择0.5作为局部损失的系数。

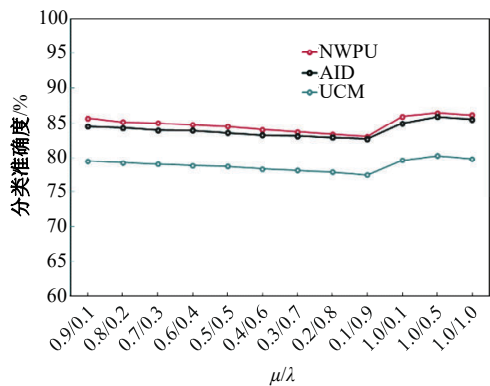


图 7 不同加权系数分类准确度对比

Fig. 7 Comparison of classification accuracy with different weighting factors

2.7 图像编码器对比实验

为了验证改进的图像编码器对 CLTP-ML 框架的效果,本文研究了基于多模态学习模型的不同图像编码器的性能。首先,选取的图像编码器包括 ResNet12^[15]、ResNet18^[15]和改进的图像编码器 MRNet。图 8 显示了分别使用上述 3 个图像编码器在 3 个数据集的 5-way 5-shot 任务下的分类准确度。从图中可以看出,在 NWPU 数据集中,MRNet 的分类准确度相较于 ResNet12 网络高出 1.10%。在 AID 数据集中,MRNet 的分类准确度最高,比第二名的 ResNet18 网络高出

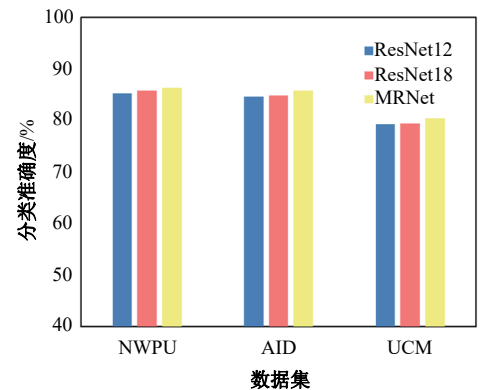


图 8 不同图像编码器分类准确度对比

Fig. 8 Comparison of classification accuracy of different image encoders

0.92%。在 UCM 数据集中,MRNet 的分类准确度能够达到 80.32%,高出 ResNet12 网络 1.13%。这有效地证明 MRNet 网络更加适合于 CLTP-ML 框架。

2.8 注意力机制对比实验

为了证明提出的新的注意力机制 ECSAM 对特征学习的有效性,本文将 ECSAM 与一些常见的注意力机制进行比较,如 SENet^[20]、ECANet^[21]和 CBAM^[22]。表 3 为 4 种注意力机制在 3 个数据集上的 5-way 5-shot 分类准确度结果。根据结果可以看出,在 NWPU 和 UCM 数据集中,ECSAM 的准确度最高,其次是 CBAM,SENet 的准确度相对最低。在 AID 数据集中,ECSAM 的分类准确度最高,相比第二名 SENet 的准确度高了 0.47%。因此,本文提出的新的注意力机制方法通过结合通道注意力机制和空间注意力机制,可以同时考虑特征通道和空间位置之间的关系,使模型能自适应地选择重要的特征通道和空间位置,从而提高模型的准确性。

表 3 不同注意力机制的分类准确度比较

Table3 Comparison of classification accuracy of different attentional mechanisms

方法	数据集	5-way 5-shot/%
SENet	NWPU	85.71±0.63
ECANet	NWPU	85.84±0.28
CBAM	NWPU	86.19±0.56
ECSAM	NWPU	86.46±0.43
SENet	AID	85.42±0.67
ECANet	AID	85.04±0.31
CBAM	AID	84.88±0.45
ECSAM	AID	85.89±0.24
SENet	UCM	79.37±0.72
ECANet	UCM	79.65±0.39
CBAM	UCM	79.97±0.24
ECSAM	UCM	80.32±0.36

2.9 可视化实验

此外,为了更直观地显示 CLTP-ML 框架中对关键特征信息提取的有效性,本文根据 XGrad-CAM^[23]方法绘制类热力图。图 9 给出了 4 组不同数据集中的图像,第一行代表原图像,第二行代表使用 FEAT 方法绘制出的类热力图,第三行代表使用 TeAw 方法绘制出的类热力图,第四行代表 CLTP-ML 框架绘制出的类热力图。图中红色区域为特征信息贡献度高的区域,蓝色区域为特征信息贡献度低的区域。对于飞机的识别,CLTP-

ML 框架将重点放在机翼与机身连接处,而 TeAw 方法将整架飞机都设为重点。对于湖泊的识别,CLTP-ML 框架将重点放在湖泊边缘形状突出的位置;而 FEAT 方法和 TeAw 方法将重点识别在非湖泊关键位置。对于宫殿的识别,CLTP-ML 框架将重点放在宫殿屋顶镂空部分;而 FEAT 方法和 TeAw 方法将重点放在宫殿屋顶边缘位置,但是只能重点识别到一个边缘。对于船只的识别,CLTP-ML 框架将重点放在船身的位置;而 FEAT 方法将非船只部分设为重点。因此,通过类热力图的对比实验可以看出 CLTP-ML 框架能够有效捕捉到区分类的特征。

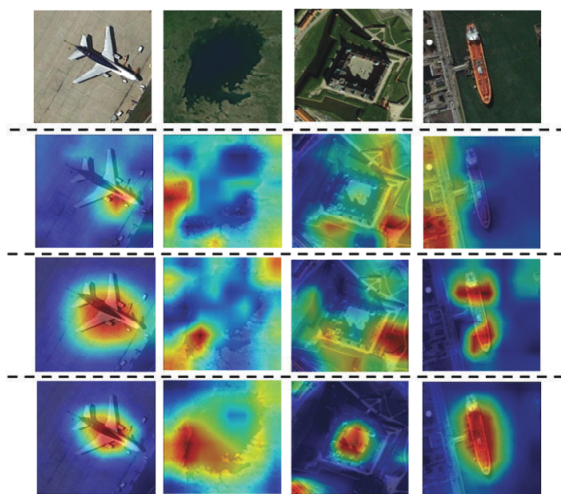


图 9 类热力图

Fig. 9 Cluster heatmap

2.10 消融实验

为了进一步探讨预训练阶段以及注意力机制对遥感图像分类的影响,本文进行了消融实验。表 4 给出了在 3 个遥感数据集上是否进行预训练阶段(P)、是否采用文本信息融合(T)以及注意力机制(A)在 5-way 5-shot 任务下的分类准确度结果。从实验结果中看出,在 NWPU 数据集中,加入预训练阶段比不加入预训练阶段的分类准确度高出 2.45%,使用文本信息融合后分类准确度提升了 0.21%,加入注意力机制模块后,分类准确度又提升了 0.41%。在 AID 数据集中,不加入预训练阶段的分类准确度可以达到 82.19%,而加入预训练模块和文本信息融合后可以提高 3.09%,再加入注意力机制模块后分类准确度可以达到 85.89%。在 UCM 数据集中加入预训练阶段的分类准确度相较于不加入预训练阶段的分类准确度提高了 1.63%,在此基础上,加入文本

信息融合后准确度又提高 0.84%,再加入注意力机制模块后,分类准确度达到了 80.32%。因此,在 3 个数据集中加入预训练阶段、加入文本信息和注意力机制可以使模型的分类准确度提高,使得模型能够更加有效地学到区分类的特征。

表 4 消融实验

Table 4 Ablation experiment

方法	数据集	分类准确度/%
CLTP-ML	NWPU	83.39±0.63
CLTP-ML+P	NWPU	85.84±0.28
CLTP-ML+P+T	NWPU	86.05±0.58
CLTP-ML+P+T+A	NWPU	86.46±0.43
CLTP-ML	AID	82.19±0.89
CLTP-ML+P	AID	84.83±0.24
CLTP-ML+P+T	AID	85.28±0.55
CLTP-ML+P+T+A	AID	85.89±0.36
CLTP-ML	UCM	77.26±0.72
CLTP-ML+P	UCM	78.89±0.52
CLTP-ML+P+T	UCM	79.73±0.44
CLTP-ML+P+T+A	UCM	80.32±0.41

3 结束语

本文提出了一种基于对比学习文本感知的小样本遥感图像分类框架 CLTP-ML。在模型中,通过预训练阶段进行全局和局部损失优化图像编码器,从而提高分类精度;通过引入文本特征进一步提高模型的感知能力;通过改进图像编码器结构使其更加适用于遥感图像分类任务。除此之外,通过设计一个新的注意力机制方法使模型可以自适应地选择重要的特征通道和空间位置,使模型能够更大程度捕捉到区分类的特征。模型分别在 3 个遥感数据集上进行了对比实验,实验结果表明:CLTP-ML 框架与经典小样本遥感分类框架相比分类结果更准确,同时也说明了使用多模态学习解决同种类别相似度低的问题比单一模态更加有效。

参考文献:

- [1] Li L, Han J, Yao X, et al. DLA-MatchNet for few-shot remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59 (9): 7844-7853.
- [2] Li X, Shi D, Diao X, et al. SCL-MLNet: boosting few-shot remote sensing scene classification via self-supervised contrastive learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-12.

- [3] 崔璐, 张鹏, 车进. 基于深度神经网络的遥感图像分类算法综述[J]. 计算机科学, 2018, 45(6): 50-53.
Cui Lu, Zhang Peng, Che Jin. A review of deep neural network based remote sensing image classification algorithms[J]. Computer Science, 2018, 45(6): 50-53.
- [4] Gong T, Zheng X, Lu X. Meta self-supervised learning for distribution shifted few-shot scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.
- [5] Yuan Z, Tang C, Yang A, et al. Few-shot remote sensing image scene classification based on metric learning and local descriptors[J]. Remote Sensing, 2023, 15(3): No. 831.
- [6] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[J]. Advances in Neural Information Processing Systems, 2017, 30: No. 05175.
- [7] Sung F, Yang Y, Zhang L, et al. Learning to compare: relation network for few-shot learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1199-1208.
- [8] Ye H J, Hu H, Zhan D C, et al. Few-shot learning via embedding adaptation with set-to-set functions [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 8808-8817.
- [9] Cheng G, Cai L, Lang C, et al. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-11.
- [10] Cheng K, Yang C, Fan Z, et al. TeAw: Text-aware few-shot remote sensing image scene classification [C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Rhodes Island, Greece, 2023: 1-5.
- [11] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.
- [12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: No. 03762.
- [13] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning, Jeju Island, Republic of Korea, 2021: 8748-8763.
- [14] Li Y, Zhu Z, Yu J-G, et al. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(12): 10590-10603.
- [15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778.
- [16] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: an all-mlp architecture for vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 24261-24272.
- [17] Cheng G, Han J, Lu X. Remote sensing image scene classification: benchmark and state of the art[J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.
- [18] Xia G S, Hu J, Hu F, et al. AID: a benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.
- [19] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification[C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, USA, 2010: 270-279.
- [20] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7132-7141.
- [21] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Online, 2020: 11534-11542.
- [22] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV), Glasgow, England, 2018: 3-19.
- [23] Fu R, Hu Q, Dong X, et al. Axiom-based gradcam: Towards accurate visualization and explanation of cnns[J/OL]. [2023-10-11]. <https://arXiv preprint arXiv:2008.02312>, 2020