

基于双注意力机制的零样本建筑图像分类方法

宁园园,张素兰,陈飞

(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘要:零样本建筑图像分类是在标记训练样本不足以涵盖所有类的情况下,利用已知建筑类别与未知建筑类别之间的知识迁移对未知类样本进行分类。针对建筑风格分类中标记数据少及局部判别性特征定位不准确的问题,提出一种基于双注意力机制的零样本图像分类方法。该方法首先引入通道注意和空间注意两种模型以增强图像特定区域的表示。其中,通道注意网络学习不同通道权重以定位图像中的建筑物;空间注意网络将位置信息嵌入通道注意图捕获目标中的细节特征,获取具有通道和空间双层维度的特征表示。其次,为减少空间映射过程中出现的信息损失,使用生成器重建视觉特征。最后,设计公共空间嵌入的零样本建筑图像分类模型,在子空间对齐视觉特征和语义特征,通过最近邻匹配实现分类任务。实验结果表明,所提方法较当前零样本学习方法而言,在零样本数据集 CUB 及建筑风格数据集 Architecture Style Dataset 上的平均分类准确率分别提高 1.3 和 0.7 个百分点。

关键词:建筑风格分类;零样本学习;双注意力机制;通道注意力;空间注意力;空间映射

中图分类号:TP183

文献标识码:A

文章编号:1673-629X(2023)10-0035-07

doi:10.3969/j.issn.1673-629X.2023.10.006

Zero-shot Architectural Image Classification Method Based on Dual Attention Mechanism

NING Yuan-yuan,ZHANG Su-lan,CHEN Fei

(School of Computer Science and Technology,Taiyuan University of Science and Technology,
Taiyuan 030024,China)

Abstract:Zero-shot architectural image classification is to use the knowledge transfer between known architectural categories and unknown architectural categories to classify unknown class samples when the labeled training samples are not enough to cover all classes. Aiming at the problems of less labeled data and inaccurate localization of local discriminative features in architectural style classification, zero-shot image classification method based on dual attention mechanism is proposed. Firstly, two models of channel attention and spatial attention are introduced to enhance the representation of specific regions of the image. Among them, the channel attention network learns different channel weights to locate the buildings in the image, the spatial attention network embeds the location information into the channel attention map to capture the detailed features in the target, and obtains feature representations with two dimensions of channel and space. Secondly, to reduce the loss of information during the spatial mapping process, a generator is used to reconstruct visual features. Finally, a zero-shot architectural image classification model with common space embeddings is designed to align visual features and semantic features in subspace and implemented the classification task through nearest neighbor matching. The experimental results show that compared with the current zero-shot learning method, the proposed method improves the average classification accuracy by 1.3 and 0.7 percentage points on the zero-shot dataset CUB and Architecture Style Dataset, respectively.

Key words:architectural style classification; zero-shot learning; dual attention mechanism; channel attention; spatial attention; space mapping

0 引言

建筑风格从地理位置、安全因素、建筑材料等方面考虑,每种风格依赖于建筑元素的结构,且都有其独特

的特征表示。例如:玫瑰窗是哥特式风格独有的特征;巴洛克风格的主要特征是圆顶、圆拱门等。建筑风格分类旨在通过建筑元素以及建筑元素之间的空间关系

收稿日期:2022-11-26

修回日期:2023-03-28

基金项目:山西省自然科学基金(202103021224285);太原科技大学研究生教育创新项目(SY2022062)

作者简介:宁园园(1997-),女,硕士研究生,研究方向为图像处理与计算机视觉;通信作者:张素兰(1971-),女,教授,博士,CCF高级会员(66965M),研究方向为数据挖掘与计算机视觉。

预测建筑的风格类别,准确的分类对建筑历史研究、建筑遗产保护和城市建设方面都具有重要意义。

近年来,建筑风格分类已取得一些重要的研究成果。Xu 等人^[1]提出在多项式潜在逻辑回归(Multinomial Latent Logistic Regression, MLLR)中引入概率分析,解决 25 类风格的分类问题。Ren 等人^[2]设计概率层次图表示基本元素的结构,从具有一致标签的 3D 模型中训练贝叶斯网络对中国古建筑基本元素的语义属性和层次结构进行编码。Yi 等人^[3]收集 17 种建筑类别的图像及描述信息,并采用卷积神经网络模型对美国房屋风格进行分类。Yoshi-mura 等人^[4]训练深度卷积神经网络对 34 个建筑师的多个作品进行分类,通过训练网络模型的权重计算建筑的视觉相似性。然而,上述方法都需要收集大量的有标签样本,但在建筑风格分类中,不同建筑风格之间存在相似性,同一建筑风格中又存在差异性^[1],导致标注更加困难。尤其对于建筑遗产图像,因为建筑景点需要被保护,不能对外开放,如故宫中的一些殿宇,图像数据难以获得,数据集中的标签样本根本不足以涵盖所有类别。因此,在缺少足够训练数据的情况下,如何利用已知建筑风格实例对未知建筑图像风格进行分类成为一个难点。

零样本分类技术旨在对训练阶段未出现过的样本类别进行分类,该技术根据已知类和未知类之间的语义相关性,将已知类的知识迁移用于未知类的识别,可有效解决样本标签缺乏时的分类问题。目前零样本学习应用于计算机视觉、自然语言处理等领域。如图 1 所示,针对建筑图像标记数据少甚至某些类别没有标注数据的情况,在建筑风格分类任务上使用零样本分类技术,缓解各风格样本分布不均衡导致的识别率低下的问题,进一步提高建筑图像分类精度。

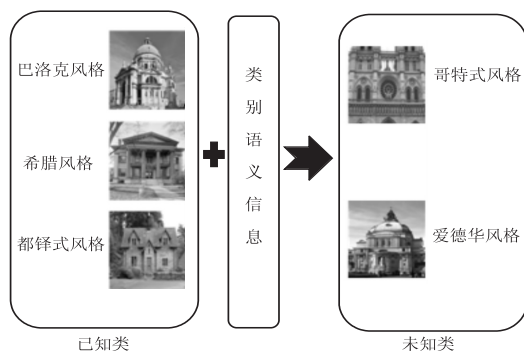


图 1 零样本学习示意图

零样本学习技术的关键是学习一个嵌入空间,根据嵌入空间的不同主要分为语义空间嵌入、视觉空间嵌入、公共子空间嵌入。语义空间嵌入是将图像特征映射到语义空间中,度量与语义描述向量的匹配度,匹配度最高的类标签为测试类输入图像的标签。视觉空

间嵌入将语义向量映射到视觉特征空间中来保留更多描述信息,能够从一定程度上缓解语义空间嵌入的枢纽点问题。但直接学习视觉空间和语义空间之间的映射函数,导致模型泛化能力较弱,影响分类性能。而公共子空间嵌入^[5]充分利用视觉和语义两种模态信息的互补性和一致性,将视觉特征和语义特征映射到公共子空间中,能够有效缓解域偏移问题。但由于在建筑图像分类任务中,每种建筑外观特征整体相似,细节元素存在差异,如哥特式建筑从上到下由尖顶、玫瑰窗、飞扶檐、尖拱门组成,而巴洛克由圆拱门、穹顶构成。从空间组成上看每个建筑元素对分类任务的重要程度不同,若采用传统的卷积神经网络,以最后一个卷积层的特征作为特征表示,则可能缺乏对建筑结构的针对性,忽略图像的各个通道和每个空间位置的重要程度,很难提取到鲁棒性较高的元素特征。

视觉注意力能够注意到与任务相关的区域,提取更有鉴别性的视觉特征。常见的注意力机制有通道注意力、空间注意力、时间注意力等。其中,通道注意力学习通道的权值并进行交互,而空间注意力通过嵌入位置信息,学习空间中重要的区域。将通道注意力与空间注意力组成的混合注意力网络学习图像特征各个维度的权重,并通过特征加权可捕获图片不同物体不同位置的细节特征。因此,针对建筑图像标签缺失及局部判别性区域定位不准确的问题,提出一种基于双注意力机制的零样本建筑图像分类方法。通过通道注意力网络自适应学习每个通道权重,选择图像中建筑物本身,忽略背景噪声影响;使用空间注意力对特征图每个位置生成掩码并加权输出,提取与分类任务相关的细节特征。同时,在学习各空间的映射中,采用生成器对映射后的特征重建,缓解空间映射过程中的信息损失问题,以保留更多原始信息,进而提高建筑图像分类精度。

1 相关工作

1.1 零样本学习

Larochelle 等人^[6]在 2008 年首次为解决字符分类问题提出了零样本学习。当前,零样本图像分类应用于图像标注、跨模态检索、目标检测等领域。根据嵌入空间的不同,零样本图像分类可分为语义空间嵌入、视觉空间嵌入、公共子空间嵌入。Ding 等人^[7]利用边缘去噪策略和自适应图训练潜在语义编码器生成潜在语义表示,提高视觉-语义映射函数的泛化。但由于语义特征映射的维度较大,容易出现枢纽点问题,使将多个类别原型的近邻点误分类。为缓解枢纽点问题,保留更多语义描述信息,提出将语义向量映射到视觉空间。Zhang 等人^[8]提出视觉空间嵌入,结合多种语义

模式进行多模态特征融合并以端到端方式联合优化。由于零样本分类中类别的视觉特征和语义特征在空间中的流形分布不同,且空间之间的维度相差较大,直接学习不同空间的映射会导致知识迁移能力较差。若通过学习一个公共子空间,实现视觉特征和语义特征对齐,可增强模型的泛化能力。赵鹏等人^[9]根据已知类的视觉特征以及类别语义之间的关系,构建了未知类的视觉特征,学习所有类别的视觉特征和语义特征到子空间的映射,并通过编码-解码器重构技术缓解了知识迁移过程中遇到的域偏移和信息丢失问题。

1.2 建筑风格分类

目前的建筑风格分类方法大多采用监督学习方法。Chen 等人^[10]通过使用一个集成的卷积神经网络模型作为全局分类器建立了建筑标注图像数据集 (Annotated Image Database of Architecture, AIDA) 并生成场景类和建筑类别的预测标签。Obeso 等人^[11]提出使用网络输入处的稀疏特征以及原色像素值对墨西哥建筑物的图像进行分类。Shalunts 等人^[12]使用局部特征的聚类寻找窗户的梯度方向,从而根据窗户的几何规则对不同建筑风格的类型进行分类,但该方法没有考虑其他建筑元素对建筑风格的影响,而且数据收集具有局限性。为缓解类别数据量不均衡的问题,Zhao 等人^[13]设计基于 GoogleNet 的深度神经网络,对数据集的数量进行增强,提高建筑风格分类性能。Chu 等人^[14]提出模拟空间配置提取可视化模型,解决目标建筑的缩放、旋转和变形问题,扩充小类别样本的数量。总之,这些方法一般需要大量标注样本,对没有标记样本的类别如何分类研究甚少。

1.3 注意力机制

注意力机制能够从无关的背景区域中提取出具有

重要信息的目标区域,目前已成功应用于视频分类、传统图像分类、机器翻译和场景分割等方面。Hou 等人^[15]将空间坐标信息整合到生成的通道注意力的特征向量中,避免全局池化造成位置信息损失,精准地定位和识别感兴趣的目标。Li 等人^[16]提出了将通道注意力和空间注意力结合的方法,使模型聚焦于关键信息,并利用注意增强技术使模型捕获特定于类的区域,提高遥感图像的分类性能。考虑到图像中不同建筑元素以及元素细节为风格分类任务贡献的权重不同,导致在提取图像特征时无法对特征进行区分,该文将通道注意力和空间注意力融合嵌入神经网络学习中,获得图像不同元素中细节位置的权重值,进而定位到判别性区域。

2 文中方法

2.1 定义

在零样本建筑风格分类任务中,已知类样本被定义为 $s = \{x_i^s, y_i^s\}$, $x_i^s \in X^s$, $y_i^s \in Y^s$, s 代表已知类, x_i^s 代表已知类的第 i 个图像, y_i^s 代表第 i 个图像的类标签。未知类样本被定义为 $u = \{x_j^u, y_j^u\}$, $x_j^u \in X^u$, $y_j^u \in Y^u$, u 代表未知类, x_j^u 代表未知类的第 j 个图像, y_j^u 代表第 j 个图像的类标签, X 和 Y 代表建筑图像集合和风格类标签集合。已知类和未知类标签不相交,即 $Y^s \cap Y^u = \emptyset$ 。用户定义属性集合 $A = \{a_1, a_2, \dots, a_n\}$, $n = i + j$, a_i 为属性向量, n 为类别总数。

2.2 双注意力机制的零样本建筑图像分类模型 (Dual Attention Mechanism for Zero-Shot Learning, DAM-ZSL)

文中分类模型由特征提取、属性编码和空间映射与分类模块组成,主要框架如图 2 所示。

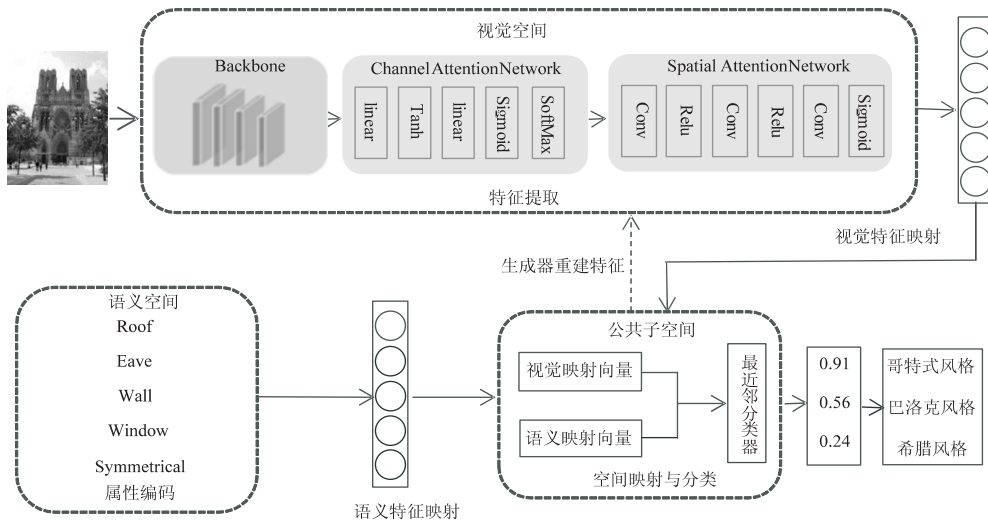


图 2 双注意力机制的零样本建筑图像分类模型

2.2.1 特征提取

视觉判别性特征提取由主干网络、通道注意网络

和空间注意网络组成。主干网络 (Backbone) 提取图像的全局特征。通道注意网络 (Channel Attention

Network, CAN) 去除图像中的天空、人、车等无关建筑的元素,定位图像中重要的建筑主体。空间注意网络 (Spatial Attention Network, SAN) 提取具有空间信息的特征表示,将建筑主体中对分类任务影响更大的建筑元素赋予更高的权重。

该文使用 ResNeXt 残差网络作为图像特征提取器,ResNeXt 作为 ResNet 的升级版,使用了 ResNet 的重复层策略及 GoogleNet 的分裂转换合并 (split-transform-merge) 的思想。在相同的参数数量下,ResNeXt 提取的特征有更强的表示能力,使图像分类的精度更高。ResNeXt-101 的每层是由多个 ResNeXt 块组成,如表 1 所示,在使用 ResNeXt-101 网络提取特征时,需要去除最后的全连接层和池化层,只保留特征提取部分。

表 1 ResNeXt-101 的网络结构

层数	ResNeXt-101
Conv1	Conv-size=7×7, stride=2 (conv+bn+relu+maxpooling)
Layer 1	Block ×3
Layer 2	Block ×4
Layer 3	Block ×23
Layer 4	Block ×3

由于通过 ResNeXt-101 提取的特征图有 2 048 个通道,使得判别特征分散。为了更好地进行建筑主体性特征定位,首先使用 1×1 卷积压缩通道 $F_{\text{tmp}} = \text{conv}_{1 \times 1}(f_{\text{ResNeXt}})$,在不改变特征图中建筑元素空间信息的情况下删除冗余通道。通道注意网络如图 2 所示,利用全局平均池化 (GAP) 计算各个通道特征图的特征值 V (公式 1),再计算各个通道的注意权值 W_{channel} (公式 2),通过 SoftMax 使每个权重的和等于 1。将通道权值作用于全局特征图上,得到通道注意图 F_{ca} (公式 3)。

$$V = \text{GAP}(F_{\text{tmp}}), V \in R^{1 \times 1 \times 16} \quad (1)$$

$$W_{\text{channel}} = \text{SoftMax}(f_{\text{c_attention}}(F_{\text{tmp}})), W_{\text{channel}} \in R \quad (2)$$

$$F_{\text{ca}} = F_{\text{tmp}} \times W_{\text{channel}}, F_{\text{ca}} \in R^{8 \times 8 \times 16} \quad (3)$$

建筑图像中并不是所有的区域都同等重要,只有与属性标签相关的建筑元素才是需要关注的,如玫瑰窗、尖拱门等是哥特式风格需要关注的空间区域。空间注意网络就是寻找建筑图像空间中重要的区域进行处理。为了突出重要像素,一些研究者使用递归神经网络计算图像的空间权值,但这种方法会将特征展开为一维向量,破坏建筑元素的空间结构,为避免空间结构的破坏,使用多层卷积组成的空间注意模型来提取像素点的空间权值 W_{spatial} (公式 4),将空间权值应用于通道特征图上计算注意特征图 F_{att} (公式 5)。

$$W_{\text{spatial}} = f_{\text{s_attention}}(F_{\text{ca}}), W_{\text{spatial}} \in R^{8 \times 8} \quad (4)$$

$$F_{\text{att}} = F_{\text{ca}} \times W_{\text{spatial}}, F_{\text{att}} \in R^{8 \times 8 \times 16} \quad (5)$$

2.2.2 属性编码

语义特征由建筑图像的语义属性构成,表示各风格类别之间的关系,是零样本建筑风格分类的关键信息。利用 one-hot 对所有类别的属性特征进行编码,0 代表无该属性,1 代表有该属性。如将哥特式建筑的属性 [rose - windows, narrow - window, glazing, ..., Symmetrical, curve, gable] 编码为 [1, 0, 1, ..., 1, 1, 0]。

2.2.3 空间映射与分类

视觉特征由双注意力得到的特征图构成,保留建筑图像中与属性相关的判别性信息。为更好地对齐视觉特征和语义特征,通过全连接层将建筑图像的视觉特征映射到公共子空间中,视觉特征到公共子空间中映射函数为 $\varphi(x_i) = W_1 \times f_{\text{att}}$ 。在学习视觉特征到子空间的映射时,由于每层的下采样操作,使得包含未知类别的判别信息损失,因此使用生成器对特征进行重建,从而减少信息的损失。同时将所有类别编码后的属性向量也通过全连接层映射到同一子空间中,学习语义映射函数 $\varphi(a_j) = W_2 \times a_j$ 。

在映射的语义向量中利用最近邻算法寻找与训练集的视觉特征相匹配的向量,预测样本的类别标签,即 $y(x) = \text{argmin} D(\varphi(x_i), \varphi(a_j))$, D 代表距离度量函数,文中使用欧氏距离作为度量函数。

2.3 模型优化

为更好地优化模型,该文使用特征重建损失、中心损失、回归损失和交叉熵损失来训练 DAM-ZSL 模型。

将视觉特征映射到公共子空间时,由于维度差异,导致一些与属性相关的判别信息在知识迁移过程中丢失,为减少信息损失,提出使用生成器对映射后的特征进行重建,计算重建损失 (公式 6), $\varphi^{-1}(\varphi(x))$ 是生成器重建后的视觉特征向量。

$$l_r = \|F_{\text{att}} - \varphi^{-1}(\varphi(x))\|_2^2 \quad (6)$$

通过最小化重建损失,使重建的视觉特征更接近实际数据,以缓解特征映射过程中的信息损失问题。

在零样本学习的训练任务中,视觉嵌入函数将视觉特征映射到公共子空间中,学习已知类的类原型特征 C_k ,即视觉特征的平均向量 (公式 7), m 为每个类别的样本总数。数据集中存在类内差异大,类间差异小的特点,因此使用中心损失函数缩小类内距离 (公式 8),将类别相同的样本更紧凑。

$$C_k = \frac{1}{m} \sum_{i=1}^m \varphi(x_i) \quad (7)$$

$$l_c = \frac{1}{m} \|\varphi(x_i) - C_k\|_2^2 \quad (8)$$

为了使嵌入的视觉特征与相应的语义属性嵌入向量接近,使用回归损失 (公式 9) 来最小化嵌入向量之

间的误差。

$$l_{\text{reg}} = \|\varphi(x_i^c) - \varphi(a^c)\|_2^2 \quad (9)$$

在分类任务中,常使用交叉熵损失(公式 10)计算预测和真实标签之间的损失值。

$$l_{\text{ce}} = \frac{1}{N} \sum_i \log \frac{\exp(\langle \varphi(x), \varphi(a) \rangle)}{\sum_c \exp(\langle \varphi(x), \varphi(a_c) \rangle)} \quad (10)$$

$c \in y_s$

因此,该文总的损失函数为(公式 11):

$$l = l_{\text{ce}} + l_r + l_c + l_{\text{reg}} \quad (11)$$

基于双注意力机制的零样本分类的目标函数为(公式 12):

$$\min_{W_1, W_2} \varphi(W_1, X) + \varphi(W_2, A) + \varphi^{-1}(W_1^T, W_1 * X) \quad (12)$$

2.4 整体算法

基于双注意力机制的零样本建筑图像分类具体流程如下:

算法 1 DAM-ZSL 算法

输入:已知类样本集合 $s = \{x_i^s, y_i^s\}$

属性标签集合 $A = \{a_1, a_2, \dots, a_n\}$, 迭代次数 n 为 200

输出:视觉映射矩阵 W_1 , 语义映射矩阵 W_2

Step 1:将图像 x_i^s 初始化为 256×256 , 输入特征提取网络,并提取图像的全局特征 f_{tmp}

Step 2:提取具有通道和空间信息的注意特征图

for $i = 1$ to n

(1) $W_{\text{channel}} = \text{SoftMax}(f_{\text{c_attention}}(F_{\text{tmp}}))$, $W_{\text{channel}} \in R$ 求通道权值

(2) 利用 $F_{\text{ca}} = F_{\text{tmp}} \times W_{\text{channel}}$, $F_{\text{ca}} \in R^{8 \times 8 \times 16}$ 求通道注意图

(3) 利用 $W_{\text{spatial}} = f_{\text{s_attention}}(F_{\text{ca}})$, $W_{\text{spatial}} \in R^{8 \times 8}$ 求空间权重

(4) 利用 $F_{\text{att}} = F_{\text{ca}} \times W_{\text{spatial}}$, $F_{\text{att}} \in R^{8 \times 8 \times 16}$ 提取包含通道和空间信息的注意特征图 F_{att}

//根据公式(6)特征重建损失训练通道-空间注意网络

End for

Step 3:利用 one-hot 对所有类别的属性特征进行编码

Step 4:空间映射:初始化 W_1, W_2

for $i = 1$ to n

(1) 利用 $\varphi(x_i) = W_1 \times f_{\text{att}}$ 计算视觉映射矩阵 W_1

(2) 利用 $\varphi(a_j) = W_2 \times a_j$ 计算语义映射矩阵 W_2

//根据公式(11)的中心损失、回归损失以及交叉熵损失函数训练网络

End for

End

3 实验分析

3.1 数据集

该文提出的模型在具有代表性的零样本数据集 CUB-200-2011 (CUB)^[17] 以及建筑风格数据集 Architecture Style Dataset^[1] 上作评估。数据集描述如表 2。CUB 共有 11 788 张图片,有 312 个类别属性,包

括 200 个鸟类别,其中 150 个类别作为已知类,50 个类别作为未知类。

表 2 数据集描述

	CUB	Architecture Style Dataset
训练集	150 类	20 类
测试集	50 类	5 类
语义属性	312 维	31 维

传统的建筑风格分类是依据标记的类标签对图像进行分类,没有考虑到风格的语义属性对分类精度的提高,根据 Yi 等人^[3]提出的美国房屋风格类别的属性特征以及建筑领域的专业知识,该文在 Architecture Style Dataset^[1] 数据集中增加类别的语义属性,属性维度为 31 维,共有 5 000 张建筑图像,包含 25 个风格类别,其中训练集 20 个类别共 4 042 张图像,测试集有 5 个类别共 958 张图像,如表 2。

3.2 实验细节

该文使用 ResNeXt-101 作为图像特征提取器,将数据集的图像大小初始化为 256×256 , 因此 ResNeXt-101 的最后一个卷积特征图的大小为 $2\ 048 \times 8 \times 8$ 。同 TransZero^[18] 一样,使用 SGD 优化器(动量为 0.9,衰减率为 0.000 1)对模型进行优化,设置 batch 大小为 50,迭代 200 次来训练模型,并设置学习率为 0.000 1。

3.3 基准实验对比模型

该文采用所有未知类的平均 top-1 精度,即对所有未知类正确预测的均值(Average Class Accuracy, ACA)(公式 13)作为评价分类的标准:

$$ACA = \frac{1}{|y|} \sum_{c=1}^{|y|} \frac{\text{第 } c \text{ 类中正确预测样本数}}{\text{第 } c \text{ 类所有样本数}} \quad (13)$$

为更好地将 DAM-ZSL 与其他先进的模型(如 DAP^[19], ALE^[20], AREN^[21], APN^[22], LDF^[23], TransZero^[18], LsrGAN^[24] 等)作比较,该文分别在零样本通用数据集和建筑风格数据集上进行实验。

如表 3 所示,在通用数据集 CUB 上,DAM-ZSL 模型的平均精度为 75%,比 TransZero 模型提高了 1.3 个百分点,说明 DAM-ZSL 能够学习与属性信息高度相关的视觉特征表示,学习的视觉区域更能表现图像的主体对象。

在表 3 中,将零样本学习用于 Architecture Style Dataset 上,DAM-ZSL 模型分类精度为 39.1%,相比 TransZero 模型和 AREN 模型分别提高 0.7 个百分点和 0.9 个百分点,表明将通道和空间注意力应用于零样本分类模型中能够聚焦图像的细节元素区域。同时,将 DAM-ZSL 与 LsrGAN 算法进行比较,分类精度提高了 0.2 个百分点,说明虽然 LsrGAN 算法利用语义正则化损失(Semantic Regularized Loss)使生成的未知类图像更加接近真实图像,但由于没有充分考虑建筑图像中与

语义向量相关的细节元素的视觉特征对分类任务的影响,从而导致其精度略低于 DAM-ZSL 模型。与 APN 属性原型网络相比,文中模型分类结果稍差,原因在于 APN 学习属性原型,将属性原型定位到视觉区域中,能够更加有效地减少匹配样本数量,在数据量小的数据集中影响更大。

表 3 不同模型方法在两个数据集上的比较

模型方法	分类精度 ACA/%	
	CUB	Architecture Style Dataset
DAP	40	—
ALE	54.9	—
AREN	71.8	38.2
APN	72	39.8
LDF	67.5	36.8
TransZero	73.7	38.4
LsrGAN	60.3	38.9
Ours	75	39.1

图 3 展示建筑风格数据集中 5 个类别的预测值与真实值之间的混淆矩阵。可以看出哥特式风格的准确率较高,原因在于其自身的建筑元素与其他未知类别的建筑元素相差较大,如玫瑰窗是其独有的,不会导致误分类。帕拉迪奥式建筑的整体对称是对已知类中古罗马和希腊建筑对称性的传承,能够学习到帕拉迪奥式建筑与已知类之间的语义属性关系,实现语义迁移,但其十字拱与巴洛克建筑的圆拱门存在语义干扰,容易导致其准确率稍差。

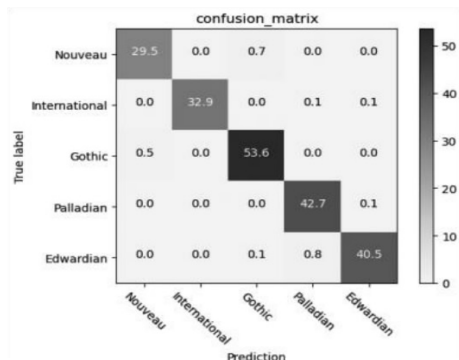


图 3 Architecture Style Dataset 未知类的混淆矩阵(%)

3.4 消融实验

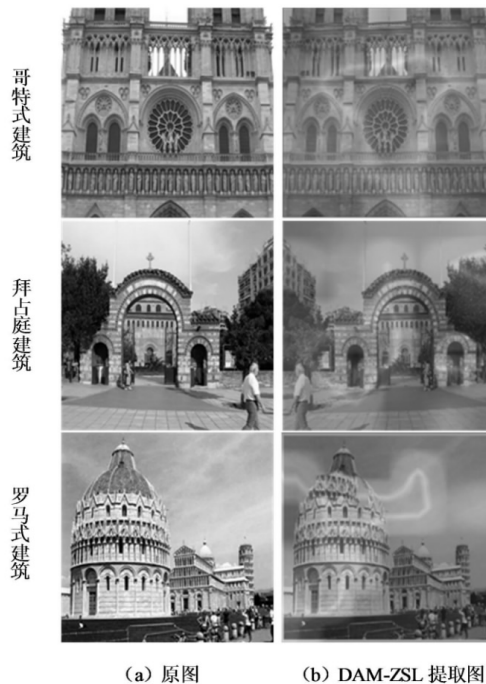
为进一步评估双注意力机制对零样本分类结果的影响,进行了消融实验,结果如表 4 所示。当不使用双注意力机制时,分类精度明显小于完整的模型(DAM-ZSL)。在 CUB 数据集中,精度下降 1.2 百分点,在 Architecture Style Dataset 中,下降了 2.6 百分点。当使用通道注意力时,由于数据集中收集的图像参差不齐, CUB 数据集中图像的目标主体更清晰,而建筑风格数据集中图像包含建筑主体及树、人、车等非建筑元素,因此分类结果对于 CUB 数据集影响不大,但对建筑风

格小数据集上提高 1.8 百分点,说明通道注意力模型能够有效地去除背景等非建筑元素的影响,提取到建筑物本身。当使用空间注意力而不使用通道注意力时,提取到的视觉特征是图像中空间结构性较强的区域,对建筑风格数据集来说,图像的空间组成较明显,因此分类精度提高 2.1 百分点。当结合通道注意力和空间注意力对图像提取特征时,能提取图像中与属性相关的视觉区域,使分类结果更准确。

表 4 双注意力机制对 ACA 精度的影响 %

通道注意力	空间注意力	CUB	Architecture Style Dataset
×	×	73.8	36.5
√	×	73.6	38.3
×	×	74.1	38.6
√	√	75.0	39.1

为了直观地表示双注意力机制在提取局部判别特征的有效性,使用 Grad-CAM 将 DAM-ZSL 模型提取出的注意特征图可视化,如图 4 所示。文中模型能够提取出与建筑风格分类相关的局部细节特征,如哥特式建筑的玫瑰窗,这说明将通道-空间双注意力网络引入零样本分类任务中使学习到的视觉特征更加具有判别性。



(a) 原图 (b) DAM-ZSL 提取图

图 4 注意图可视化

4 结束语

为了对训练集中未知类的建筑图像进行正确分类,提出了一种基于双注意力机制的零样本建筑图像分类方法,结合通道注意机制和空间注意机制提取了建筑风格图像中与属性相关的判别性特征,同时将局

部判别性特征和属性特征映射到同一子空间中,使公共子空间中存在丰富的视觉信息和类别语义属性之间的关系信息,并使用最近邻算法实现了对未知建筑风格样本的有效分类。在之后的工作中将根据语义信息结合图卷积网络构建类别之间的关系,进一步提高零样本建筑风格图像分类结果。

参考文献:

- [1] XU Z, TAO D, ZHANG Y, et al. Architectural style classification using multinomial latent logistic regression[C]//European conference on computer vision. Switzerland: Springer, 2014:600–615.
- [2] REN P, ZHOU M, WANG Z, et al. A probabilistic model for traditional Chinese architecture[C]//2016 international conference on virtual reality and visualization (ICVRV). Hangzhou: IEEE, 2016:411–417.
- [3] YI Y K, ZHANG Y, MYUNG J. House style recognition using deep convolutional neural network[J]. Automation in Construction, 2020, 118(5):1–15.
- [4] YOSHIMURA Y, CAI B, WANG Z, et al. Deep learning architect: classification for architectural design through the eye of artificial intelligence[J]. Computational Urban Planning and Management for Smart Cities, 2019, 9(8):249–265.
- [5] 秦牧轩, 荆晓远, 吴飞. 基于公共空间嵌入的端到端深度零样本学习[J]. 计算机技术与发展, 2018, 28(11):44–47.
- [6] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]//Proceedings of the AAAI conference on artificial intelligence. Chicago: AAAI, 2008:646–651.
- [7] DING Z, LIU H. Marginalized latent semantic encoder for zero-shot learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019:6191–6199.
- [8] ZHANG L, XIANG T, GONG S. Learning a deep embedding model for zero-shot learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017:2021–2030.
- [9] 赵鹏, 汪纯燕, 张思颖, 等. 一种基于融合重构的子空间学习的零样本图像分类方法[J]. 计算机学报, 2021, 44(2):409–421.
- [10] CHEN J, STOUFFS R, BILJECKI F, et al. Hierarchical (multi-label) architectural image recognition and classification[C]//Proceedings of the 26th international conference of the association for computer-aided architectural design research in Asia. Hong Kong: CAADRIA, 2021:161–170.
- [11] OBESO A M, BENOIS-PINEAU J, ACOSTA A Á R, et al. Architectural style classification of Mexican historical buildings using deep convolutional neural networks and sparse features[J]. Journal of Electronic Imaging, 2016, 26(1):11–16.
- [12] SHALUNTS G, HAXHIMUSA Y, SABLATNIG R. Architectural style classification of building facade windows[C]//International symposium on visual computing. Berlin: Springer, 2011:280–289.
- [13] ZHAO P, MIAO Q, LIU R, et al. Architectural style classification based on DNN model[C]//Chinese conference on pattern recognition and computer vision (PRCV). Xi'an: Springer, 2019:505–516.
- [14] CHU W T, TSAI M H. Visual pattern discovery for architecture image classification and product image search[C]//Proceedings of the 2nd ACM international conference on multimedia retrieval. Hong Kong: Association for Computing Machinery, 2012:1–8.
- [15] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville: IEEE, 2021:13713–13722.
- [16] LI F, FENG R, HAN W, et al. An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13(7):3862–3878.
- [17] WAH C, BRANSON S, WELINDER P, et al. The caltechucsd birds-200–2011 dataset[J]. California Institute of Technology, 2011, 12(7):1–8.
- [18] CHEN S, HONG Z, LIU Y, et al. Transzero: attribute-guided transformer for zero-shot learning[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI, 2022:330–338.
- [19] AKATA Z, PERRONNIN F, HARCHAOUI Z, et al. Label-embedding for attribute-based classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Portland: IEEE, 2013:819–826.
- [20] AKATA Z, PERRONNIN F, HARCHAOUI Z, et al. Label-embedding for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(7):1425–1438.
- [21] XIE G S, LIU L, JIN X, et al. Attentive region embedding network for zero-shot learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019:9384–9393.
- [22] XU W, XIAN Y, WANG J, et al. Attribute prototype network for zero-shot learning[J]. Advances in Neural Information Processing Systems, 2020, 33(12):21969–21980.
- [23] LI Y, ZHANG J, ZHANG J, et al. Discriminative learning of latent features for zero-shot recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018:7463–7471.
- [24] VYAS M R, VENKATESWARA H, PANCHANATHAN S. Leveraging seen and unseen semantic relationships for generative zero-shot learning[C]//European conference on computer vision (ECCV), Glasgow: Springer, 2020:70–86.