计算机科学
COMPUTER SCIENCE

**基于多级特征融合与注意力模块的场景识别方法**

许华杰，秦远卓，杨洋

引用本文

许华杰，秦远卓，杨洋. 基于多级特征融合与注意力模块的场景识别方法[J]. 计算机科学, 2022, 49(4): 209-214.

XU Hua-jie, QIN Yuan-zhuo, YANG Yang. Scene Recognition Method Based on Multi-level Feature Fusion and Attention Module[J]. Computer Science, 2022, 49(4): 209-214.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100135

# 基于多级特征融合与注意力模块的场景识别方法

许华杰[1,2]　秦远卓[1]　杨　洋[1]
1 广西大学计算机与电子信息学院　南宁 530004
2 广西多媒体通信与网络技术重点实验室　南宁 530004
(hjxu2009@163.com)

**摘　要**　场景图像通常由背景信息和前景目标对象构成,用于场景识别任务的卷积神经网络(CNN)通常需要根据场景中关键目标的特征,甚至结合目标之间的位置关系来识别出场景所属类别。针对场景图像中较小尺寸的关键目标特征随着网络层次的加深而逐渐消失,从而导致场景识别错误的问题,提出了一种基于多级特征融合与注意力模块的场景识别方法。首先,将深度神经网络 ResNet-18 的特征提取部分划分出 5 个分支;然后,将 5 个分支输出的多级特征进行融合,利用融合后的特征进行场景识别和分类,以弥补丢失的目标信息;最后,在网络中加入改进的注意力模块,以达到着重学习场景图像中关键目标的目的,进一步提升识别效果。在多个场景数据集上进行实验对比,结果表明,所提方法在 MIT-67,SUN-397 和 UIUC-Sports 这 3 个场景数据集上的识别准确率分别达到了 88.2%,79.9%和 97.7%,相比目前主流的场景识别方法其具有更高的识别准确率。

**关键词**:场景识别;卷积神经网络;特征融合;注意力模块
**中图法分类号**　TP391

## Scene Recognition Method Based on Multi-level Feature Fusion and Attention Module

XU Hua-jie[1,2],QIN Yuan-zhuo[1] and YANG Yang[1]
1 College of Computer and Electronic Information,Guangxi University,Nanning 530004,China
2 Guangxi Key Laboratory of Multimedia Communications and Network Technology,Nanning 530004,China

**Abstract**　Scene image is usually composed of background information and foreground objects. Convolutional neural network (CNN) used for scene recognition task usually needs to recognize the category of scene according to the characteristics of key objects in the scene,or even combined with the position relationship between objects. Aiming at the problem that the key target features of small size in the scene image gradually disappear with the deepening of the network level,which leads to scene recognition errors,a scene recognition method based on multi-level feature fusion and attention module is proposed. Firstly,the feature extraction part of the deep neural network ResNet-18 is divided into five branches,and then the multi-level features of the output of the five branches are fused,and the fused features are used for scene recognition and classification to make up for the lost target information. Secondly,an improved attention module is added to the network to achieve the purpose of focusing on learning the key targets in the scene image,so as to improve the recognition effect further. Experimental results on several scene datasets show that the recognition accuracy of the proposed method on MIT-67,SUN-397 and UIUC-Sports scene datasets reaches 88.2%,79.9% and 97.7% respectively,which is higher than the current mainstream scene recognition methods.

**Keywords**　Scene recognition,Convolutional neural network,Feature fusion,Attention module

## 1　引言

场景识别是计算机视觉领域的基本任务之一,其研究目标是使计算机能够对图像进行处理,自动识别和理解图像中的场景信息,它在自动驾驶、道路交通、机器人和视频监控等应用领域都发挥着十分重要的作用。近年来,卷积神经网络(Convolutional Neural Network,CNN)在计算机视觉领域取得了巨大成功,研究人员开始应用 CNN 来解决图像场景识别问题[1-2]。文献[1]和文献[3]分别对场景图像分类的相关技术和场景识别中的深度学习方法进行了深入的调研和综述,介绍了相关技术和方法的主要研究内容和发展情况。较早的基于 CNN 的研究都只利用了 CNN 模型的高层特征进行场景识别,并没有明确使用中间层的特征。

场景识别任务不同于传统的目标识别,场景通常是由

到稿日期:2021-01-18　返修日期:2021-05-20
基金项目:广西壮族自治区科技计划项目(2017AB15008);崇左市科技计划项目(FB2018001)
This work was supported by the Science and Technology Plan Project of Guangxi Zhuang Autonomous Region(2017AB15008) and Science and Technology Plan Project of Chongzuo(FB2018001).
通信作者:杨洋(520012399@qq.com)

多个目标组成的，也就是说，在识别一个场景时，网络需要检测到用于“支持”该场景的多个“目标对象”（如图画、画笔和绘图板共同构成“绘画室”场景）作为判断依据，某些场景的识别还需要考虑目标之间的位置关系。研究过程中发现，许多目标尺寸较小的场景图像识别错误率较高，文献[4]指出，这是由于 CNN 中的下采样操作使小尺寸目标的特征信息在网络的较深层中变得不明显甚至消失，导致分类器利用高层特征进行识别时缺少了关键信息，容易将其误判为其他相似的场景。最新的研究指出，对于小尺寸目标信息在 CNN 高层特征中缺失，导致场景识别错误的问题，可以利用多级特征融合的思路来解决[5]。

此外，还有研究发现，场景所属类别通常与图像中的某些目标联系更紧密，如健身房场景中的健身器械、机房场景中成排的电脑、玩具商店场景中的玩具等，这些目标可以看作场景识别的关键目标，在提取特征时，着重学习场景的关键目标信息，有利于提升识别准确率[6]。Woo 等提出的卷积块注意力模块 CBAM（Convolutional Block Attention Module）[7] 在目标检测等任务中可以有效捕捉关键目标信息，并且可融合到各种 CNN 中进行端到端的训练。

基于上述讨论，本文提出了一种以 ResNet-18 网络为基础架构的改进场景识别方法。本文的主要工作分为以下两部分：

(1)提出了一种用于场景识别任务的多级特征融合网络（Multi-level CNN）。通过将 ResNet-18 网络分为 5 个分支，融合 5 个分支输出的特征形成新的特征表示并用于最终的识别，以弥补 CNN 高层特征中小尺寸目标信息的缺失。

(2)在 CBAM 的基础上提出改进的注意力模块 S-CBAM，并将 S-CBAM 融入到多级特征融合网络中，使网络能将注意力集中到重要的通道和空间上，从而着重学习场景中关键目标的特征信息，进一步提升场景的识别准确率。

## 2 相关工作

### 2.1 基于多级特征融合的场景识别方法

将 CNN 多层特征进行融合是一种有效改进场景识别方法的思路[3]。相比只利用高层特征的方法，多级融合可以将场景图像的多级信息进行综合，实现特征互补，有利于达到更好的识别效果。文献[8]使用在 Places 数据集上预训练的 CNN 模型来提取场景图像特征，合并最后两个全连接层的输出来表示图像。文献[9]使用在 Places 数据集上预训练的 CNN 模型来提取场景图像特征，并基于特征融合和特征选择提出了 RF-CNN 模型。文献[5]提出了一种多层集成网络来提高关键目标比较小的场景的识别率，在多个低层特征后增加分类器进行单独识别，然后在网络中进行集成学习来作最终识别。文献[10]针对 GoogleNet 的网络结构，根据辅助损失函数将其分为 3 个部分并将得到的卷积特征进行融合，提出了 G-MS2F 模型。文献[11]提出了一种局部卷积监督层（LCS），通过绕过 CNN 中的一个卷积层并直接连接到最终损失函数来增强局部卷积特性，并用 Fisher 卷积矢量对局部信息进行编码，与全连接层特征相结合构成 LS-DHM 表示。采用类似思想的还有特征金字塔网络（Feature Pyramid Networks，FPN）[12]，其通过提取多尺度的特征信息并加以融合，来提高目标检测的精度，特别是对于小物体检测的精度。FPN 还可以与经典网络组合来提升原网络效果。

多级特征融合需要充分考虑网络的结构特性，根据不同 CNN 的特点来提取并选择特征进行融合，同时还要考虑到融合模型的参数规模和计算复杂度。

### 2.2 注意力模块

当前应用较广、较为典型的注意力模块有 SENet[13] 和 CBAM[7]。SENet 的出现是为了解决在卷积池化过程中，特征图不同通道的重要性不同引发的信息损失问题，其更注重特征的通道信息。卷积块注意力模块 CBAM 比 SENet 多了空间注意力模块，其效果相对更好，同时可添加到卷积神经网络中的卷积层后面，因此其通用性也更好。在许多分类或检测模型中加入 CBAM 模块可以使性能得到不同程度的提升[7]，根据处理对象特征和网络结构特性对 CBAM 进行合理改进也是提升性能的有效途径之一。

## 3 方法描述

### 3.1 多级特征融合网络

#### 3.1.1 ResNet-18

ResNet-18 是何恺明团队提出的残差网络 ResNet[14] 中的一种，这里的 18 指带权重的层为 18 层（包括卷积层和全连接层，不包括池化层和 BN 层），具体由 1 个卷积核大小为 7×7 的卷积层（Conv）、1 个内核大小为 3×3 的池化层（最大池化 Max Pooling）、4 组残差模块（Residual Module）和 1 个全连接层（FC）构成，每组残差模块内有 2 个标准残差块结构（Residual Block）。

将 ResNet-18 提取的特征图进行可视化以了解 ResNet-18 对场景图像的特征学习过程，如图 1 所示，从左至右、从上至下分别展示了输入图像和层次由浅至深提取到的部分特征图样例。通过观察分析特征图可以看到，网络的低层特征包含更多的线条、纹理、目标形状等信息，随着层次深度的增加，图像信息逐渐减少，一些关键目标特征也越来越模糊甚至消失。CNN 的低层特征由于来源的层次较浅，因此学习到的特征更多地指向场景图像的局部和细节信息，而来源于更深层次的高层特征，则更强调图像全局结构及主要目标的整体轮廓等抽象信息，因此高层特征适合用于物体分类识别任务[5]。由于场景图像相比一般的物体图像含有更多复杂纹理、关键目标和空间关系，图像的细节信息在 CNN 深层逐渐消失，导致高层特征缺少了可区分性的场景识别依据，从而产生可能识别错误的问题。

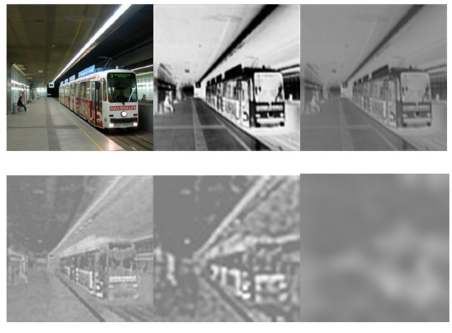


图 1 ResNet-18 不同层次的特征图

Fig. 1 Feature maps from different layers of ResNet-18

### 3.1.2 多级特征融合网络

根据上述对 ResNet-18 提取的特征图可视化结果以及可能引发场景识别错误的原因分析，本文在 ResNet-18 网络的基础上进行改进，提出了更适合场景识别任务的包含 1 个主网络和 5 个分支的多级特征融合网络（Multi-level CNN），其结构如图 2 所示。

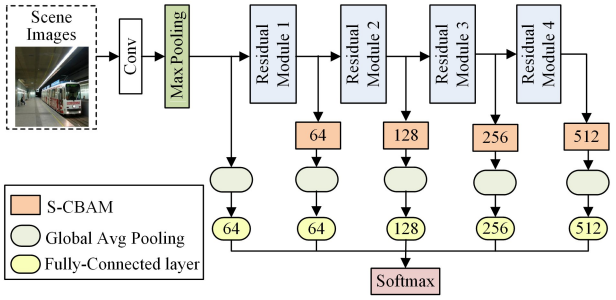



图 2 多级特征融合网络

Fig. 2 Multi-level feature fusion network

ResNet-18 包含 8 个残差模块，残差模块是 ResNet 网络的核心结构，是影响网络特征提取效果的关键，因此，在划分阶段要保证每个残差模块的完整性并以残差模块作为分支插入边界。由于相邻模块生成的特征有一定的相似性，将每个模块的特征均进行融合不仅会带来信息冗余，还会增加网络的复杂度和计算量，因此考虑根据下采样操作将 ResNet-18 的特征提取部分划分为 5 个阶段。之所以选择 5 个阶段，是因为经过实验对比发现，综合考虑识别效果和计算效率两方面因素，采取 5 个阶段融合是相对较优的方案（相比其他方案而言，如 4 个阶段或 6 个阶段融合的方案）。第 1 阶段包含 1 个卷积层和 1 个池化层，第 2—5 阶段均由 2 个残差模块组成，这 5 个阶段形成 Multi-level CNN 的 5 个分支。首先将这 5 个分支输出的特征向量分别经过全局平均池化层（Global Avg Pooling）和全连接层（fully-conneced layer），各分支的全连接层设置的神经元数与该分支输入特征的通道数保持一致，分别为 64，64，128，256 和 512（详见图 2，各分支的通道数标注在相应的方框内），其目的在于对各分支的输入特征进行学习并降低特征维度使其满足合并要求。然后将所得的 5 个特征向量在通道维度上进行串联合并，形成一个 1 024 维的特征向量。最后将组合后的特征向量输入到 SoftMax 分类器中进行预测，得到识别的结果。第 $i$ 个输入图像的融合特征向量 $V(i)$ 如式（1）所示：

$$V(i)=\bigcup_{k=1}^{5}\{fc(global_{\text{avg}}(v^k(i)))\} \tag{1}$$

其中，$v^k$ 表示第 $k$ 个分支的输出特征向量，$global_{\text{avg}}(\cdot)$ 为全局平均函数，$fc(\cdot)$ 为全连接操作，$\cup$ 代表串联合并（concatenation）操作。

## 3.2 场景卷积块注意力模块

CBAM 由通道注意力子模块 CA 和空间注意力子模块 SA 组成。通道注意力描述了“关键对象是什么”的问题，而空间注意力描述了“关键对象在哪里”的问题。本文借鉴 CBAM 的整体结构和注意力权重计算方法，并在其基础上提出了场景卷积块注意力模块（Scene CBAM，S-CBAM），其核心思想是根据场景识别任务的特点，将 CBAM 的 CA 和 SA 分别改进为场景通道注意力子模块 SCA 和场景空间注意力子模块 SSA；此外，根据 Multi-level CNN 的结构特点，将 S-CBAM 插入 4 个残差模块和第 2—5 分支之间，并设置 S-CBAM 的通道维数与该分支输入特征的通道维数保持一致，分别为 64，128，256 和 512（详见图 2），使网络能够对关键目标投入更多关注，进一步提升识别准确率。

S-CBAM 以 Multi-level CNN 的 4 个残差模块的输出特征作为输入 Input，并将其依次与 SCA 和 SSA 计算得到的注意力权重相乘后再与自身相加得到模块输出结果 Output，其整体结构如图 3 所示。

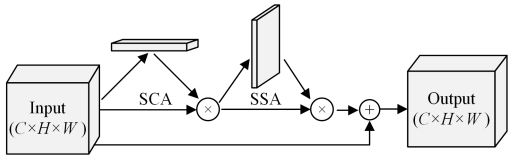



图 3 场景卷积块注意力模块

Fig. 3 Scene CBAM

### 3.2.1 场景通道注意力子模块（SCA）

通道注意力需要通过全局池化将各通道上的特征图都压缩为一个实数来作为该通道的权重，再施加到对应的特征图上。全局平均池化可以聚合特征图的空间信息，全局最大池化（Global Max Pool）有助于聚合不同目标的突出特征。

通过实验发现，CBAM 中通道注意力模块对输入特征分别进行全局平均池化和全局最大池化计算后再相加的做法并不比只采用其中一种池化取得的识别效果更好，由此推测，前者的处理方式没有很好地将两种池化的优势相结合，导致其在场景图识别中无法充分发挥两种池化方法的作用。因此，为了更好地结合两种池化的优势，对通道注意力模块进行改进：在改进后的通道注意力模块 SCA 中，分别进行两种池化运算并将结果串联合并，再利用多层感知机结构（MLP）学习两者的内在联系。改进后的通道注意力模块（SCA）的结构如图 4 所示。

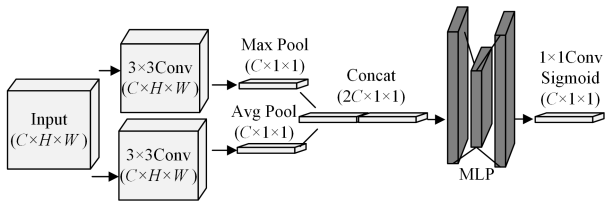



图 4 改进的通道注意力模块

Fig. 4 Scene Channel Attention

SCA 对大小为 $C\times H\times W$ 的输入特征分别进行全局最大池化和全局平均池化，将得到的结果进行串联合并（concat）后输入 MLP，该结构两端的全连接层神经元数为 $2C$，中间隐层的神经元数为 $C$，最后通过 $1\times1$ 卷积将特征尺寸降维至 $C\times1\times1$，利用 Sigmoid 激活函数将结果映射到（0，1），以获得标准的特征通道注意力权重 $M_{\text{SCA}}$，其计算式如下：

$$M_{SCA}=\sigma(W(FC(global_{avg}(Input)\cup global_{max}(Input))))\tag{2}$$

其中，$global_{avg}(\cdot)$为全局平均函数，$global_{max}(\cdot)$为全局最大函数，两个函数输出结果的特征尺寸均为$C\times1\times1$，$\cup$代表串联合并操作，$FC(\cdot)$为 MLP 结构，$W(\cdot)$为卷积运算，$\sigma(\cdot)$为 Sigmoid 激活函数。

3.2.2 场景空间注意力子模块(SSA)

空间注意力需要对输入特征进行空间位置上的全局池化操作，将特征图每个像素点位置的一维特征压缩为一个实数作为空间权重，再施加到对应的特征图位置上。

通过实验发现，将 CBAM 的空间注意力模型单独作用于场景图像识别的提升效果有限，因此，本文根据场景识别的特点设计了新的空间注意力模块 SSA。原 CBAM 的空间注意力子模块对输入特征分别进行全局最大池化和全局平均池化计算后，再将计算结果合并，但实验发现，在场景识别任务中使用单独的全局最大池化的效果实际上优于全局平均池化及两种池化的结合，因此 SSA 采取单独的全局最大池化进行运算。考虑到场景图像中的关键目标尺寸大小不一致这一特点，在 SSA 中加入了多卷积核的组合架构，使其可以学习到不同尺度感受野的特征信息，以适应不同尺度的目标对象，改进后的空间注意力模块(SSA)的结构如图 5 所示。

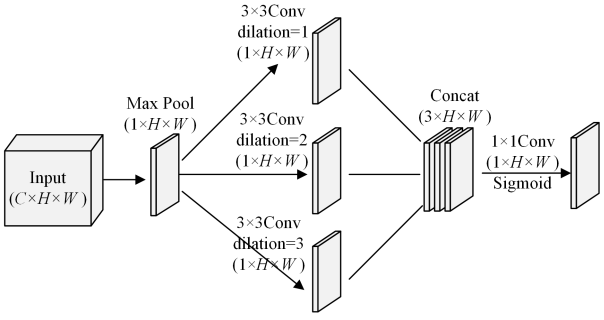



图 5 改进的空间注意力模块

Fig. 5 Scene spatial attention

SSA 对尺寸为$C\times H\times W$的输入特征在空间位置上进行最大池化(Max Pool)操作得到$1\times H\times W$的空间特征，空间特征分别经过 3 个 3×3 空洞卷积，它们的扩张系数分别为 1,2,3，再将 3 个卷积的结果串联合并，最后通过 1×1 卷积降维至$1\times H\times W$，利用 Sigmoid 激活函数将结果映射到(0,1)，以获得标准的特征空间注意力权重$M_{SSA}$，其计算公式如式(3)、式(4)所示：

$$C_i=W_i(MaxPool(Input)),i=1,2,3\tag{3}$$

$$M_{SSA}=\sigma(W(C_1\cup C_2\cup C_3))\tag{4}$$

其中，$MaxPool(\cdot)$为空间位置上的全局最大池化函数，沿着通道轴对特征点求最大值；3 种卷积操作符分别为$W_1$，$W_2$，$W_3$，其输出结果分别为$C_1$，$C_2$，$C_3$；$\cup$代表串联合并操作；$W(\cdot)$为卷积运算；$\sigma(\cdot)$为 Sigmoid 激活函数。

## 4 实验

### 4.1 实验数据集和相关参数设置

实验采用 MIT-67，SUN-397 和 UIUC-Sports 这 3 个实验数据集。MIT-67[15]是一个专门面向室内场景识别的数据集，共包含 67 类室内场景，总样本数为 15 620；SUN-397[16]是面向视觉识别任务的一个大规模通用数据集，包含室外场景、自然场景和室内场景 3 个大类下的 397 个场景类别，每个类别的样本数超过 100；UIUC-Sports[17]是一个体育活动场景数据集，涵盖 8 类场景。实验中 3 个实验数据集均按 8∶2 的比例划分为训练集和测试集，每组实验重复 3 次，取 3 次的平均值作为最终实验结果。

根据数据集特点及相关经验，主要实验参数设置如下：输入图像尺寸归一为 224 * 224，批处理大小设置为 100，使用随机梯度下降(SGD)优化方法，设定初始学习率为 0.01，动量为 0.9，权重衰减率为 0.0005。实验环境如下：Windows10 64 位操作系统，四核 Intel Core i7 CPU(2.7 GHz)，NVIDIA GeForce GTX1080Ti 显卡(GPU)，16GB 内存。实验在 PyTorch 框架下实现。

### 4.2 实验及其结果

4.2.1 多级特征融合效果测试

通过比较 Multi-level CNN 模型(以下简称 M-CNN)和 ResNet-18 模型在这 3 个场景数据集上的识别率来测试多级特征融合的效果，结果如表 1 所列。

表 1 M-CNN 与 ResNet-18 模型之间的对比

Table 1 Comparisons between M-CNN and ResNet-18 models

| Model | MIT-67 | SUN-397 | UIUC-Sports |
|---|---|---|---|
| ResNet-18 | 74.6 | 63.6 | 89.2 |
| M-CNN | **79.3** | **72.9** | **91.7** |

由表 1 可看出，M-CNN 模型在 MIT-67，SUN-397 和 UIUC-Sports 这 3 个场景数据集上的识别率均高于 ResNet-18 模型，分别高出 4.7，9.3 和 2.5 个百分点。

为了更直观地展示 M-CNN 和 ResNet-18 模型的对比效果，给出两个网络模型对场景数据集样本的预测分数(prediction score)，预测分数描述的是模型对样本识别的结果与其真实标签相符的概率，分数越高表示识别结果正确的可能性就越高。由于篇幅有限，仅从每个数据集中随机选择 3 个样本进行识别结果和预测分数的展示，如图 6 所示。由图 6 可看到，与 ResNet-18 相比，M-CNN 能够显著提高对场景类别识别的预测得分。例如，图 6(a)中第一张图的真实标签为“书店(bookstore)”，M-CNN 的预测得分为 0.688 2，而 ResNet-18 的预测得分为 0.143 4，因此将其错误预测为“图书馆(library)”；图 6(b)中间的图真实标签为“直升飞机场(heliport)”，M-CNN 的预测得分为 0.653 4，而 ResNet-18 仅为 0.024 0，因此将其错误识别为“荒地(badlands)”。导致上述现象的原因可能是在这些关键目标尺寸较小的场景(易错场景)图像中，ResNet-18 的高层特征很有可能将小尺寸关键目标的特征信息丢失，而 M-CNN 通过将 ResNet-18 分割成 5 个阶段，将各阶段输出的特征进行融合，有效解决了小尺寸目标的特征信息随着网络层次的加深而变得不明显或消失，从而导致识别错误的问题。

Class：bookstore
ResNet-18：0.1434
M-CNN：0.6882

Class：bakery
ResNet-18：0.4614
M-CNN：0.7827

Class：jewelry shop
ResNet-18：0.2010
M-CNN：0.6611

(a)MIT-67

Class：gymnasium
ResNet-18：0.3432
M-CNN：0.7596

Class：heliport
ResNet-18：0.0240
M-CNN：0.6534

Class：garage dump
ResNet-18：0.3205
M-CNN：0.6107

(b)SUN397

Class：bocce
ResNet-18：0.1602
M-CNN：0.5336

Class：rowing
ResNet-18：0.1221
M-CNN：0.6517

Class：snowboarding
ResNet-18：0.2244
M-CNN：0.6009

(c)UIUC-Sports

图 6 两种模型预测分数比较的示例

Fig. 6 Examples of prediction score comparisons between M-CNN and ResNet-18

### 4.2.2 注意力模块效果测试

为了测试 M-CNN 中的注意力模块 S-CBAM 的作用，对比 M-CNN 在去掉注意力模块(图 2 中右边 4 个分支中都去掉 S-CBAM 模块，只包含全局平均池化层和全连接层)、加入 CBAM 和加入 S-CBAM 这 3 种情况下对场景数据集的识别准确率，从而测试改进的注意力模块 S-CBAM 的效果，实验结果如表 2 所列。

表 2 注意力模块识别准确率的比较

Table 2 Comparisons of recognition accuracies of attention modules

| Model | MIT-67 | SUN-397 | UIUC-Sports |
|---|---|---|---|
| M-CNN (No S-CBAM) | 79.3 | 72.9 | 91.7 |
| M-CNN+CBAM | 84.5 | 76.4 | 96.4 |
| M-CNN+S-CBAM | **88.2** | **79.9** | **97.7** |

对比表 2 的结果可知，在 M-CNN 中加入注意力模块可以有效提升其识别率，进一步对比后两行分别对应加入 CBAM 和 S-CBAM 的实验结果可以发现，加入改进后的 S-CBAM 在 3 个数据集上的识别率比加入 CBAM 分别高出 3.7%，3.5% 和 1.3%，这验证了改进后的 S-CBAM 的有效性。导致这一结果的原因有 3 个：1)S-CBAM 的通道注意力模块(SCA)将全局最大池化和全局平均池化的输出特征串联合并，再利用多层感知机学习合并特征的做法，保留了更多特征信息，同时更好地结合了全局平均池化和全局最大池化的优点；2)S-CBAM 的空间注意力模块(SSA)加入的多卷积核组合结构可以更好地学习场景图像的多尺度信息，适合场景图像中关键目标大小不一的情况，其空间权重分配机制更合理；3)S-CBAM 将 SCA 和 SSA 相结合，综合两者的优点，因此其在场景识别上取得了更好的表现。

### 4.2.3 场景识别方法对比测试

为了测试本文所提方法的性能，分别在 MIT-67，SUN-397 和 UIUC-Sports 这 3 个场景数据集上将提出的方法与其他主流的场景识别方法进行对比，3 个数据集上的识别准确率对比结果分别如表 3－表 5 所列。

表 3 MIT-67 数据集上的识别准确率对比

Table 3 Comparisons of the recognition accuracies on MIT-67 datasets

(单位：%)

| Model | MIT-67 |
|---|---|
| RF-CNNs[9] | 72.4 |
| G-MS2F[10] | 79.6 |
| MVML-LSTM[8] | 80.5 |
| Bai 等[18] | 80.8 |
| CS(VGG-19)[19] | 82.2 |
| LS-DHM[11] | 83.8 |
| Proposed | **88.2** |

表 4 SUN-397 数据集上的识别准确率对比

Table 4 Comparisons of the recognition accuracies on SUN-397 datasets

(单位：%)

| Model | SUN-397 |
|---|---|
| Bai 等[18] | 59.5 |
| MVML-LSTM[8] | 63.0 |
| G-MS2F[10] | 64.1 |
| CS(VGG-19)[19] | 64.5 |
| LS-DHM[11] | 67.6 |
| Proposed | **79.9** |

表 5 UIUC-Sports 数据集上的识别准确率对比

Table 5 Comparisons of the recognition accuracies on UIUC-Sports datasets

(单位：%)

| Model | UIUC-Sports |
|---|---|
| Meng 等[20] | 83.0 |
| GOC[21] | 83.1 |
| LPR[22] | 86.5 |
| LCST-LDA-CNN[23] | 91.5 |
| RF-CNNs[9] | 94.9 |
| Proposed | **97.7** |

由表 3 可知，本文方法在 MIT-67 上的识别率为 88.2%，比 LS-DHM 方法高出 4.4%；由表 4 可知，本文方法在 SUN-397 上的识别率为 79.9%，比 LS-DHM 方法高出 12.3%；由表 5 可知，本文方法在 UIUC-Sports 上的识别率为 97.7%，比 RF-CNNs 高出 2.8%。综上可知，本文所提方法在 3 个数据集上的场景识别准确率均优于其他对比方法。该方法取得较优效果的原因有两方面：首先，多级特征融合方法可以综合利用图像的多级特征实现特征互补，相比只利用 CNN 高层特征的方法，其更大程度地保留了低级特征中的小尺寸目标信息作为识别依据，更符合场景识别任务的要求；其次，加入了改进的注意力模块后，加强了网络学习场景图像中的关键目标的能力，从而进一步提升了识别准确率。

**结束语** 针对场景图像的关键目标尺寸较小，目标特征

随着CNN层次的加深而缺失，从而导致场景识别错误的问题，本文提出了一种基于多级特征融合和注意力模块的场景识别方法。该方法中的多级特征融合网络M-CNN通过融合ResNet-18模型的5个阶段特征，实现了多级特征的优势互补，弥补了CNN高层特征中的小尺寸目标特征丢失的不足，对易错场景图像样本的真实类别预测分数有显著提升；将改进后的场景卷积块注意力模块S-CBAM添加到多级特征融合网络进行端到端训练，能够更好地根据场景识别的特点，加强通道注意力模块的两种池化操作的内在联系和空间注意力模块的多尺度学习能力，将注意力聚焦于关键目标的特征信息，进一步提升了场景识别的准确率。实验结果表明，本文所提方法在3个场景数据集上取得的识别性能均优于目前主流的场景识别方法，验证了本文方法的有效性。

本文所提方法采用的融合策略只是将多个分支特征进行简单的串联合并，会带来一定程度的信息冗余，如何有效减少信息冗余并提出更优的融合策略是下一步的研究工作。

## 参 考 文 献


[1] TIAN Y L, ZHANG W T, ZHANG Q S, et al. Review on Image Scene Classification Technology[J]. Acta Electronica Sinica, 2019, 47(4): 915-926.

[2] XU J L, LI L Y, WAN X J, et al. Indoor scene recognition method combined with target detection[J]. Computer Application, 2021, 41(3): 1-6.

[3] LI X Y, ZHU J, MA L N. Survey of Scene Recognition Methods Based on Deep Learning[J]. Computer Engineering and Applications, 2020, 56(5): 25-33.

[4] LUIS H, JIANG S, LI X. Scene Recognition with CNNs: Objects, Scales and Dataset Bias[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. 2016: 571-579.

[5] ZHANG L H, LI L Q, PAN X P, et al. Multi-level ensemble network for scene recognition[J]. Multimedia Tools and Applications, 2019, 78(19): 28209-28230.

[6] KUDUS A R, TEH C S. Design and Development of Scene Recognition and Classification Model Based on Human Preattention Visual Attention[J]. Journal of Physics: Conference Series, 2021, 1755(1): 1-12.

[7] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]// Proceedings of the 2018 European Conference on Computer Vision. 2018: 3-19.

[8] BAI S, TANG H D, AN S. Coordinate CNNs and LSTMs to categrize scene images with multi-views and multi-levels of abstraction[J]. Expert Systems with Applications, 2019, 120: 298-309.

[9] BAI S. Growing random forest on deep convolutional neural networks for scene categorization[J]. Expert Systems with Applications, 2017, 71: 279-287.

[10] TANG P, WANG H, KWONG S. G-MS2F: GoogleNet based multi-stage feature fusion of deep CNN for scene recognition [J]. IEEE Geoscience and Remote Sensing Letter, 2017, 225: 188-197.

[11] GUO S, HUANG W, QIAO Y. Locally supervised deep hybrid model for scene recognition[J]. IEEE Transactions on Image Processing, 2017, 26(2): 808-820.

[12] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.

[13] HU J, LI S, GANG S. Squeeze-and-Excitation Networks [C]// The 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA, 2018: 7132-7141.

[14] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the International Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[15] QUATTONI A, TORRALBA A. Recognizing indoor scenes [C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. 2009: 413-420.

[16] XIAO J, HAYS J, EHINGER K A, et al. SUN database: Large-scale Scene Recognition from abbey to zoo[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. 2010: 3485-3492.

[17] LI L J, LI F F. What, Where and Who? Classifying Events by Scene and Object Recognition[C]// Proceedings of the International Conference on Computer Vision and Pattern Recognition. 2007: 1-8.

[18] BAI S, TANG H. Categorizing scenes by exploring scene part information without constructing explicit models[J]. Neurocomputing, 2018(281): 160-168.

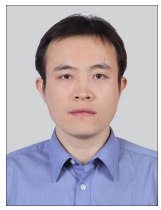
[19] XIE G S, ZHANG X Y, YAN S, et al. Hybrid CNN and dictionary-based models for scene recognition and domain adaption [J]. IEEE Transaction on Circuits & Systems for Video Technology, 2017, 27(6): 1263-1274.

[20] MENG X, WANG Z, WU L. Building global image features for scene recognition[J]. Pattern Recognition, 2012(45): 373-380.

[21] GAO C, SANG N, HUANG R. Spatial multi-scale gradient orientation consistency for place instance and scene category recognition[J]. Information Sciences, 2016(372): 84-97.

[22] SADEGHI F, TAPPEN M F. Latent pyramidal regions for recognizing scenes[C]// Proceedings of European Conference on Computer Vision. Florence, 2012: 228-241.

[23] HUANG C, LUO W, XIE Y. Local-class- shared topic latent dirichlet allocation based scene classification [J]. Multi-media Tools and Applications, 2017, 76(14): 15661-15679.

**XU Hua-jie**, born in 1974, Ph. D, associate professor, is a member of China Computer Federation. His main research interests include artificial intelligence, acoustic signal recognition and computer vision.

**YANG Yang**, born in 1995, postgraduate. Her main research interests include artificial intelligence and computer vision.

（责任编辑：柯颖）