Journal of Computer Applications
计算机应用,2021,41(9):2720－2725
ISSN 1001-9081
CODEN JYIIDU
2021-09-10
http://www.joca.cn

文章编号:1001-9081(2021)09-2720-06
DOI:10.11772/j.issn.1001-9081.2020111815

结合目标检测的室内场景识别方法

徐江浪[1],李林燕[2],万新军[1],胡伏原[1*]
(1.苏州科技大学 电子与信息工程学院,江苏 苏州 215009;
2.苏州经贸职业技术学院 信息技术学院,江苏 苏州 215009)
(*通信作者电子邮箱fuyuanhu@mail.usts.edu.cn)

摘　要:在目标检测网络(ObjectNet)和场景识别网络相结合的方法中,由于ObjectNet提取的目标特征和场景网络提取的场景特征的维度和性质不一致,且目标特征中存在影响场景判断的冗余信息,导致场景识别的准确率低。针对这个问题,提出一种改进的结合目标检测的室内场景识别方法。首先,在ObjectNet中引入类转换矩阵(CCM),将ObjectNet输出的目标特征进行转化,使得目标特征的维度与场景特征的维度相一致,以此减少特征维度不一致带来的信息丢失;然后采用上下文门控(CG)机制对特征中的冗余信息进行抑制,从而降低不相关信息的权重,提高了目标特征在场景识别中的作用。该方法在MIT Indoor67数据集上的识别准确率达到90.28%,与维护空间布局的对象语义特征(SOSF)方法相比识别准确率提高了0.77个百分点;其在SUN397数据集上识别准确率达到81.15%,与交替专家层次结构(HoAS)方法相比识别准确率提高了1.49个百分点。实验结果表明,所提方法提高了室内场景识别的准确率。

关键词:深度学习;卷积神经网络;室内场景识别;目标检测;类转换矩阵;上下文门控
中图分类号:TP391.4　　**文献标志码:**A

Indoor scene recognition method combined with object detection

XU Jianglang[1], LI Linyan[2], WAN Xinjun[1], HU Fuyuan[1*]
(1. *School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou Jiangsu* 215009, *China*;
2. *School of Information Technology, Suzhou Institute of Trade and Commerce, Suzhou Jiangsu* 215009, *China*)

Abstract: In the method of combining Object detection Network (ObjectNet) and scene recognition network, the object features extracted by the ObjectNet and the scene features extracted by the scene network are inconsistent in dimensionality and property, and there is redundant information in the object features that affects the scene judgment, resulting in low recognition accuracy of scenes. To solve this problem, an improved indoor scene recognition method combined with object detection was proposed. First, the Class Conversion Matrix (CCM) was introduced into the ObjectNet to convert the object features output by ObjectNet, so that the dimension of the object features was consistent with that of the scene features, as a result, the information loss caused by inconsistency of the feature dimensions was reduced. Then, the Context Gating (CG) mechanism was used to suppress the redundant information in the features, reducing the weight of irrelevant information, and increasing the contribution of object features in scene recognition. The recognition accuracy of the proposed method on MIT Indoor67 dataset reaches 90.28%, which is 0.77 percentage points higher than that of Spatial-layout-maintained Object Semantics Features (SOSF) method; and the recognition accuracy of the proposed method on SUN397 dataset is 81.15%, which is 1.49 percentage points higher than that of Hierarchy of Alternating Specialists (HoAS) method. Experimental results show that the proposed method improves the accuracy of indoor scene recognition.

Key words: deep learning; Convolutional Neural Network (CNN); indoor scene recognition; object detection; Class Conversion Matrix (CCM); Context Gating (CG)

0　引言

近些年来,随着互联网的普及以及人们生活水平的不断提高,场景识别技术的发展为人们带来越来越多的服务和便利。室内场景识别是场景识别的关键部分,室内场景识别技术的发展在智能家居、服务机器人、安防监控等领域都有着广阔的应用前景[1-2]。

在早期,室内场景识别一般都是利用颜色、纹理、形状等特征进行识别,随着诸如尺度不变特征变换(Scale-Invariant Feature Transform, SIFT)、加速健壮特征(Speeded-Up Robust Features, SURF)、方向梯度直方图(Histogram of Oriented Gradients, HOG)等算子的广泛应用,比较流行的分类方法是

收稿日期:2020-11-19;**修回日期:**2020-12-26;**录用日期:**2021-02-22。
基金项目:国家自然科学基金面上项目(61876121);江苏省高等学校自然科学研究项目(19KJB520054)。
作者简介:徐江浪(1996—),男,江苏宿迁人,硕士研究生,CCF会员,主要研究方向:图像处理、深度学习、场景识别;　李林燕(1983—),女,湖南岳阳人,高级工程师,硕士,主要研究方向:多模态信息处理;　万新军(1996—),男,江苏连云港人,硕士研究生,CCF会员,主要研究方向:图像处理、深度学习;　胡伏原(1978—),男,湖南岳阳人,教授,博士,CCF会员,主要研究方向:图像处理、模式识别、信息安全。

www.joca.cn

利用各种算子提取场景特征,以此训练出较好的模型实现场景识别,其中应用较广的是词袋(Bag-Of-Words, BOW)模型,产生了基于BoW模型的一系列场景识别方法。Lazebnik等[3]提出的一种基于空间金字塔(Spatial Pyramid)的BOW模型场景识别方法为后来的研究学者提供了思路,不少学者在此基础上作出了改进和创新;但是这类方法很大程度上依赖人工算子进行特征提取,导致泛化能力不足,而深度学习方法在这一方面有了很大的改善。

在基于深度学习的方法中,出现了多种以卷积神经网络(Convolutional Neural Network, CNN)为基础的方法[4-7]。由于复杂的室内场景包含多种目标,全局特征难以表达这些目标特征,基于此,Quattoni等[8]提出结合全局特征(Global Feature)和局部特征(Local Feature)来识别室内场景,利用局部特征表示目标特征,提高了识别准确率。Cheng等[9]提出一种用于场景识别的具有对象性的语义描述符方法,利用不同场景中对象配置的相关性,通过场景中所有对象的共现模式来选择有代表性的和有区别的对象,增强了类间的可区分性。Herranz等[10]提出了一种基于多尺度特征的方法,将不同尺寸的图像分别送入适合它们的目标网络和场景网络来分别提取特征,解决了图像尺寸和识别网络的匹配问题,但是尺寸的不断增加导致了方法复杂度的增加。Wang等[11]提出了一种知识指导消歧(Knowledge Guided Disambiguation)策略,利用知识网络提取的目标特征生成场景图像的软标签,指导场景网络最小化损失函数,有效地解决了类间差异小和类内差异大的问题,但是目标特征的利用率仍较低。为了使网络更加关注场景图像的显著区域,Rezanejad等[12]提出一种基于场景轮廓的场景分类方法,利用场景轮廓传达的形状和表面的几何形状作为信息输入进行场景识别,这种基于中轴的显著性度量方法增加了场景特征中有用的信息。Sun等[13]也提出一种场景识别的综合表示方法,该方法融合了从物体语义、全局外观和上下文外观这三个区别视图中提取的深层特征,利用特征的多样性和互补性,以及场景图像的上下文信息,提升了识别准确率。尽管上述方法已经取得了显著的效果,但是对于结合目标检测的室内场景识别方法而言,场景图像中目标特征的利用效率极大地影响了室内场景识别效果,而这是当前亟待解决的问题[14-17]。

由于场景图像中包含多种目标信息,但并非所有的目标都对场景识别产生积极作用,甚至有些会产生反作用影响最终的识别效果(如卧室中出现的电脑、餐厅中入镜的沙发等);而且,在进行场景特征和目标特征之间的融合时,由于两种特征的维度不一致,在特征融合时会造成信息丢失等问题。因此,本文提出了一种结合目标检测的室内场景识别方法,主要思想是:1)为了解决特征维度不一致的问题,本文利用类转换矩阵(Class Conversion Matrix, CCM)对目标特征的维度进行转换,使目标特征在与场景特征融合时能够减少信息丢失;2)针对冗余信息的问题,本文采用了上下文门控(Context Gating, CG)机制对目标特征中所包含的冗余信息进行抑制,降低不相关信息的权重,提高室内场景识别中目标特征的利用效率,使网络更加关注图像的相关目标区域,提高室内场景识别的准确度。

1　结合目标检测的室内场景识别方法

1.1　网络结构

本文以Inception为基础网络架构,网络框架由场景识别网络PlacesNet(本文图示以MR-CNNs(Multi-Resolutions CNNs)为例)、目标检测网络(Object detection Network, ObjectNet)两个部分组成,如图1所示。PlacesNet是用来提取场景特征的CNN模型;ObjectNet用来提取目标特征,为了获得更精确的目标特征,本文采用了在ImageNet上预训练过的CNN模型。然后,提取的特征进入特征融合模块(Feature Fusion Module),该模块通过卷积层+ReLU(Rectified Linear Unit)和哈达玛积的组合将两个网络的输出特征进行融合;最后,经过softmax层进行分类,输出最终结果。作为改进,本文将CCM置于ObjectNet提取的目标特征之后,对它进行转化,使它的特征维度和场景网络提取的特征维度一致;在融合目标特征和场景特征之前,采用CG对目标特征中可能存在的冗余信息进行了抑制,以此提高目标特征的利用效率。

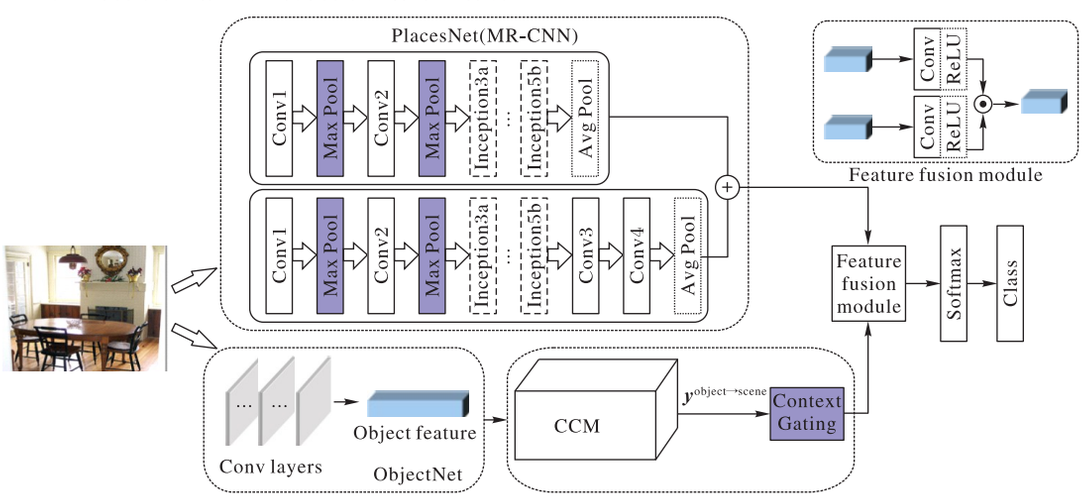


图1　网络框架

Fig. 1　Network framework

1.2　特征转换

为了将ObjectNet提取的目标特征转化为场景特征,避免目标特征与场景特征两种不同特性的特征直接进行融合,提升特征融合模块的融合效果,本文在目标检测和场景识别相结合的方法中引入CCM,如图2所示,将CCM置于ObjectNet的特征之后,将目标特征进行加权和,最终转换成和场景特征相同维度。输入目标特征$\boldsymbol{x}^{object}$是从ObjectNet中提取的,输出特征为$\boldsymbol{y}^{object\to scene}$,CCM的计算如下:

www.joca.cn

$$\boldsymbol{y}^{\text{object}\to\text{scene}} = \boldsymbol{W}\boldsymbol{x}^{\text{object}} + \boldsymbol{b} \tag{1}$$

其中：$\boldsymbol{x}^{\text{object}} \in \mathbf{R}^n$；权重 $\boldsymbol{W} \in \mathbf{R}^{m\times n}$，偏差 $\boldsymbol{b} \in \mathbf{R}^m$，$\boldsymbol{y}^{\text{object}\to\text{scene}} \in \mathbf{R}^m$ 是与 $\boldsymbol{x}^{\text{object}}$ 相关的输出。n 是输入向量的维数，m 是输出向量的维数。如果将 ImageNet 数据集用于训练对象模块，则 n=1 000。如果将 Places 2 数据集用于训练场景模块，则 m=365。CCM 不仅可以应用于不同的数据集，而且可以应用于相同的数据集，即 m=n，如具有 Places 2 数据集格式的 CCM 的参数 $\boldsymbol{W} \in \mathbf{R}^{m\times m}$ 和 $\boldsymbol{b} \in \mathbf{R}^m$，则具有相同数据集格式的输入 $\boldsymbol{x}^{\text{object}} \in \mathbf{R}^m$ 和输出 $\boldsymbol{y}^{\text{object}\to\text{scene}} \in \mathbf{R}^m$。而且，从 CCM 的权重中可以分析目标和场景之间的关系，即目标和场景之间的相关程度，如果一个目标频繁出现在场景中，权重则越高；相反则越低。这意味着场景识别性能可以获得进一步的提高。

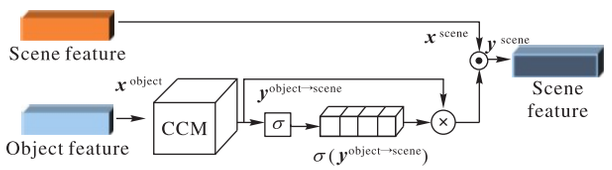


图 2 特征转换示意图

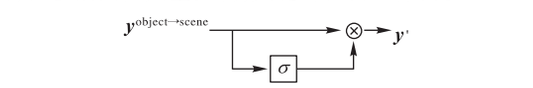
Fig. 2 Schematic diagram of feature conversion

1.3 冗余信息抑制

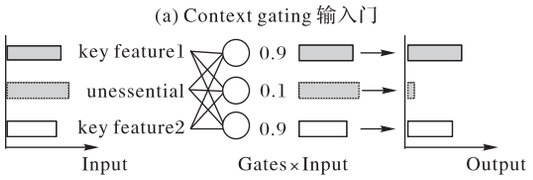
每一幅场景图像中都可能包含各种各样的目标，而当一幅场景图像中出现某一特别的对象时，这个场景就很有可能属于某一特别的类，如一张床对于卧室，或浴缸对于浴室的重要性，但是其他不相关特征可能会影响网络的判断。为了降低不相关特征的权重，本文引入了 CG，它的计算公式为：

$$\begin{aligned}\boldsymbol{y}^{'\text{scene}} &= \boldsymbol{y}^{\text{object}\to\text{scene}} \otimes \sigma(\boldsymbol{y}^{\text{object}\to\text{scene}}) = \\ &(\boldsymbol{W}\boldsymbol{x}^{\text{object}} + \boldsymbol{b}) \otimes \sigma(\boldsymbol{W}\boldsymbol{x}^{\text{object}} + \boldsymbol{b})\end{aligned} \tag{2}$$

其中：$\boldsymbol{x}^{\text{scene}}$ 表示由场景网络提取的场景特征；⊗ 表示一种矩阵相乘法；$\sigma(x) = 1/\left(1 + \exp(-x)\right)$ 是值为(0,1)的 sigmoid 激活函数。如图 3(a)，由于 sigmoid 激活函数的特性，左端无限趋于 0，右端无限趋于 1，构成一个限制信息的“门”，当不相关信息通过门控时，函数取值趋向 0，可以抑制不相关的信息。在目标特征中不相关信息有可能占据较高的视觉激活，CG 模块可以通过上述门控方法抑制这些不相关信息的传递，降低不相关特征对识别效果的影响，使网络更加关注图像的相关目标区域。如图 3(b)所示，一个场景中可能存在一个或多个关键特征，图中用 key feature1 和 key feature2 代表这类特征，不相关特征会被门控大量抑制，关键特征得以保留。



(a) Context gating 输入门



(b) Context gating 工作原理

图 3 CG 输入门及原理

Fig. 3 Input gate and working principle of CG

1.4 特征融合模块

本文采取特征融合模块用于融合来自两个网络的特征，获得最终用于分类结果预测的特征，本文采用了卷积层+ReLU 和哈达玛积的组合结构，具体如图 1 所示。场景网络获得的特征表示为：

$$\boldsymbol{y}_{\text{s}}^{\text{scene}} = \phi(\boldsymbol{W}_1\boldsymbol{x}^{\text{object}} + \boldsymbol{b}_1) \tag{3}$$

其中：$\boldsymbol{y}_{\text{s}}^{\text{scene}}$ 表示场景特征经过特征融合模块的特征表示；$\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 表示该分支的卷积层的权重和偏差；$\phi(\,)$ 表示 ReLU 激活函数。

同样，目标网络获得的特征表示为：

$$\boldsymbol{y}_{\text{o}}^{\text{scene}} = \phi(\boldsymbol{W}_2\boldsymbol{y}^{'\text{scene}} + \boldsymbol{b}_2) \tag{4}$$

其中：$\boldsymbol{y}_{\text{o}}^{\text{scene}}$ 表示目标特征经过特征融合模块的特征表示；$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 表示该分支的卷积层的权重和偏差。

最终融合之后的特征表示为：

$$\boldsymbol{y}^{\text{scene}} = \boldsymbol{y}_{\text{s}}^{\text{scene}} \odot \boldsymbol{y}_{\text{o}}^{\text{scene}} \tag{5}$$

其中：⊙ 表示哈达玛积，即矩阵对应位置相乘；$\boldsymbol{y}^{\text{scene}}$ 为经过特征融合模块的最终特征表示，其经过 softmax 层进行分类，获得最终的预测结果。

1.5 损失函数

ObjectNet 能够检测出场景图像中存在的目标对象，本文利用目标检测网络的这种优点与场景网络进行联合训练。在本文的实验中，当两个网络联合训练时，如果其中一个网络的判别性较强，整体损失较小，则会阻碍判别性较差的网络的优化。为了避免其中一个网络控制训练过程，首先，ObjectNet 和 PlacesNet 分别针对给定的场景进行训练；然后，将训练好的两个网络从头开始完全训练特征融合模块和分类器。在训练过程中，通过最小化以下目标函数来同时预测结果：

$$l = -\left(\sum_{I_i \in D}\sum_{k=1}^{m}\prod_{y_i=k}\log p_{i,k} + \lambda\sum_{I_i \in D}\sum_{k=1}^{n}\log f_{i,k}\right) \tag{6}$$

其中：D 表示训练数据集；I_i 表示第 i 张图像；y_i 表示第 i 张图像的真实标签；p_i 表示预测的场景标签；f_i 表示目标网络产生的标签；λ 是平衡两项的参数；n 和 m 分别对应两个网络的特征维度。这种联合学习方法能够通过利用包含在目标网络中的额外知识作为归纳偏差来提高泛化能力，并减少过度拟合对训练数据集的影响。图 4 为本文方法的准确率收敛曲线，在迭代 1 500 次时准确率逐渐趋于稳定。

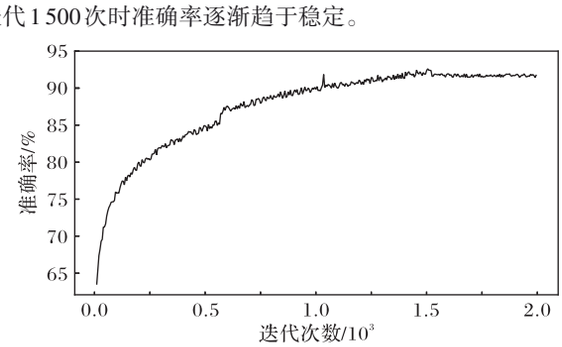


图 4 准确率收敛曲线

Fig. 4 Accuracy convergence curve

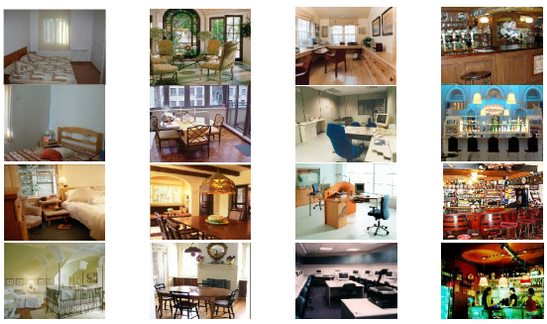
2 实验与结果分析

2.1 实验平台及数据集

本文方法的具体实现使用的是深度学习框架 Tensorflow，实验环境为 Ubuntu 16.04 操作系统，使用 4 块 NVIDIA 1080Ti 图形处理器(GPU)加速运算。本文采用数据集包括：MIT Indoor67 数据集，该数据集包含 67 个室内类别，总共 15 620 张图像，每个类别至少有 100 张图像；SUN397 数据集，它由 108 754 张图像组成，包含了 397 个图像类别；ImageNet 数据集，它是 ObjectNet 训练用来提取场景图像中的目标的，包含了 1 000 个对象类别；Scene-15 数据集，该数据集包含 15 个不

www.joca.cn

同类别的4 485个室内外场景图像,其中室内场景类别分别是卧室、工厂、厨房、客厅、办公室和商店。MIT Indoor67数据集部分场景示例如图5所示。



(a) Bedroom　(b) Dining room　(c) Office　(d) Bar

图5　MIT Indoor67数据集部分场景示例

Fig. 5　Some scene examples of MIT Indoor67 datasets

2.2　实验参数及结果分析

实验使用梯度下降法进行训练,设置权重衰减(weight dacay)系数为0.000 1,动量(momentum)系数为0.9,batch size设置为64,初始学习率设置为0.001。实验中要学习的参数是权重 $\boldsymbol{W} \in \mathbf{R}^{m \times n}$,偏差 $\boldsymbol{b} \in \mathbf{R}^{m}$,$m$ 和 n 表示维度,与训练数据集的类别数有关,因此不会增加参数的规模,对模型的学习速度不会产生大的影响。λ 是需要人为给定的参数,经过多次实验验证,当λ=0.5时,训练的效果最好。

本文使用的评价指标分别为:召回率*Rec*(Recall)、精准率*Pre*(Precision)和识别准确率*Acc*(Accuracy),其公式定义如下:

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

其中:*TP*(True Positive)表示预测为正例,实际为正例;*TN*(True Negative)表示预测为负例,实际为负例;*FP*(False Positive)表示预测为正例,实际为负例;*FN*(False Negative)表示预测为负例,实际为正例。*TP*、*TN*、*FP*和*FN*分别表示预测的真假性与实际场景的关系。本文绘制了精准率-召回率*PR*(Precision-Recall)曲线来分析模型,图6中展示了SOSF(Spatial-layout-maintained Object Semantics Features)[13]、HoAS(Hierarchy of Alternating Specialists)[18]、SDO(Semantic Descriptor with Objectness)[9]、VSAD(Vector of Semantically Aggregated Descriptors)[19]四种方法和本文方法的*PR*曲线。*PR*曲线表现了召回率和精准率之间的关系,曲线越靠近右上角,表明模型的性能越好。从图6中可以看出本文方法在这方面优于其他几种方法。

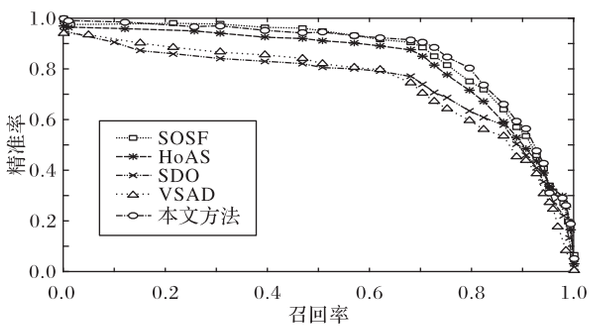


图6　不同方法的PR曲线

Fig. 6　PR curves of different algorithms

表1为本文方法在Scene-15数据集上得到的混淆矩阵以及相应的召回率和识别精度,矩阵第i行第j列的值表示第i类场景被预测为第j类场景的样本数。表1中序号分别对应Scene-15数据集中的类别,分别为:1)Living;2)MITmountain;3)MITopencountry;4)MITforest;5)store;6)MIThighway;7)CALsuburb;8)MITstreet;9)industrial;10)kitchen;11)PARoffice;12)MITinsidecity;13)MITcoast;14)bedroom和15)MITtallbuilding。从表1中可以看出,办公室的类别召回率为100.0%,精准率为95.7%;厨房的召回率也达到100.0%,卧室和客厅两个室内场景类别容易被相互误分类,分类错误的原因是客厅中的目标(如沙发)与卧室场景中的目标相似。本文方法在室外场景上也有不错的效果。

表1　Scene-15数据集的混淆矩阵

Tab. 1　Confusion matrix of Scene-15 dataset

类别	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	召回率/%
1	**190**	0	0	0	0	0	0	0	0	6	3	0	0	6	0	92.7
2	0	**267**	6	6	0	0	0	0	0	0	0	0	4	0	0	94.3
3	0	1	**191**	0	0	0	2	0	0	0	0	0	10	0	0	93.6
4	0	6	0	**215**	0	0	0	0	0	0	0	0	0	0	0	97.7
5	1	0	0	0	**211**	0	0	0	0	0	0	0	0	0	5	98.5
6	0	0	0	0	0	**158**	0	0	0	0	0	0	0	0	0	**100.0**
7	0	0	0	0	2	0	**141**	0	0	0	2	0	0	0	0	97.2
8	1	0	4	0	0	0	0	**180**	2	0	0	1	0	0	0	95.7
9	0	0	0	0	1	0	0	0	**205**	0	0	3	0	1	0	97.6
10	0	0	0	0	0	0	0	0	0	**105**	0	0	0	0	0	**100.0**
11	0	0	0	0	0	0	0	0	0	0	**114**	0	0	0	0	**100.0**
12	1	0	1	0	0	0	0	0	3	0	0	**204**	0	0	0	97.6
13	0	1	5	3	0	0	0	0	0	0	0	0	**251**	0	0	96.5
14	8	0	0	0	0	0	0	0	0	0	0	0	0	**112**	0	93.3
15	0	0	1	1	0	0	0	0	8	0	0	2	0	0	**240**	95.2
精准率/%	94.5	97.0	91.8	95.5	98.1	**100.0**	98.6	**100.0**	94.0	94.5	95.7	97.1	94.7	94.1	97.9	

www.joca.cn

2.2.1 多种方法在数据集 MIT Indoor67 和 SUN397 上的比较

目前，已有多种方法对场景识别进行了研究。本文将多种方法分别在数据集 MIT Indoor67 和 SUN397 上进行了对比实验，主要与以下几种方法作为对比实验：Object-Scale[10]、VSAD[19]、MR-CNNs（Multi-Resolution CNNs）[11]、SDO[9]、Semantic-Aware[1]、HoAS[18]、SOSF[13]，结果如表 2 所示。在 MIT Indoor67 数据集上进行的实验，VSAD[19]虽然利用了 PatchNet 来指导特征提取，但是该网络对 Local patch 的提取能力并不强，准确率只有 86.20%；MR-CNNs[11]的准确率为 86.70%，它在网络结构上进行了创新，结合了多分辨率网络，但没有利用局部特征，导致在准确率上并没有提升很多；Semantic-Aware[1]在场景的语义信息方面有所改善，但是面对复杂场景时受语义分割的影响，场景识别效果也受到限制，准确率为 87.10%；另外，HoAS[18]解决了场景的类内差异性和类间相似性的问题，SOSF[13]的结果是对比方法中效果最高的为 89.51%，但还是缺乏对场景的目标特征足够的关注。因此，本文在利用目标特征进行分析的同时，抑制了不相关特征的权重，识别准确率达到了 90.28%，相比上述方法最好的结果，仍有 0.77 个百分点的提升。同样，在 SUN397 数据集上，本文方法识别准确率达到了 81.15%，相较于另外几种方法中准确率最高的方法 HoAS（为 79.66%），准确率提高了 1.49 个百分点。实验结果表明，通过对目标特征进行转换并对冗余信息进行抑制，增强了特征的表达能力和特征融合的效果，显著提高了识别准确率。

表 2 多种方法在 MIT Indoor67 和 SUN 397 数据集上的准确率 单位：%

Tab. 2 Accuracy of various methods on MIT Indoor67 and SUN 397 datasets unit: %

方法	MIT Indoor67	SUN397
Object-Scale[10]	86.04	70.17
VSAD[19]	86.20	73.00
MR-CNNs[11]	86.70	72.00
SDO[9]	86.76	73.41
Semantic-Aware[1]	87.10	74.04
HoAS[18]	88.24	79.66
SOSF[13]	89.51	78.93
本文方法	**90.28**	**81.15**

2.2.2 目标和场景之间的关系分析

CCM 的权重体现了目标以及场景之间的关系紧密程度，本文对 CCM 的权重进行了分析，结果显示权重值越高，表明该目标越经常出现在该场景中；反之亦然。本文在 MIT Indoor67 数据集上进行实验，通过表 3~4 数据对它们的关系展开分析。表 3 和表 4 分别显示了 CCM 权重的前 3 名和后 3 名。

表 3 CCM 权重的前 3 名

Tab. 3 Top-3 of CCM weights

正序	场景	目标	权重
1	bedroom	bed	0.097 21
2	kitchen	stove	0.094 80
3	bathroom	toilet	0.093 16

表 3 数据显示的是某场景类中可能存在的合理对象，权重最高的是 bedroom 中的 bed，达到了 0.097 21，表明这种搭配合理度最高；另外两个 kitchen 和 stove 以及 bathroom 和 toilet 的权重也分别达到了 0.094 80、0.093 16，这表明它们的搭配合理性也很高。这个数据验证了 CCM 正确反映了目标对于场景的积极影响。

同理，根据表 4 的数据显示，权重值越小的值所对应的场景类别和对象类别的组合其合理性也越小，权重为负值时，表明该目标对场景识别起到了负面影响，需要对其权重进行抑制。比如表 4 中所列的 bathroom 和 stove、office 和 slipper 等组合。从实验数据可以看出，该实验结果表明了 CCM 权重对室内场景识别是有效的。本文方法合理体现了对于目标和场景之间的关系程度。

表 4 CCM 权重的后 3 名

Tab. 4 Bottom-3 of CCM weights

倒序	场景	目标	权重
1	bathroom	oven	−0.099 37
2	office	slipper	−0.096 41
3	bedroom	blackboard	−0.094 59

2.2.3 冗余信息抑制效果

为了测试本文方法对于冗余信息的抑制效果，在 MIT Indoor67 数据集上分别利用了目标检测网络和本文方法进行了效果对比。首先，利用目标检测网络检测室内场景图像中包含的目标；然后，本文以类激活图（Class Activation Map，CAM）[20]的形式分析 CG 模块对识别效果的影响，该方法以直观的视觉形式表现出了网络对于图像的关注区域。

目标特征中的冗余信息抑制结果如图 7 所示：图 7(a)为原图；图 7(b)为经过目标检测网络的输出图像，可以看出场景中一些明显的物体被检测出来，如床、桌子、椅子以及电脑等；图 7(c)为经过本文方法的 CAM 可视化效果。从图 7 可看出，相对于图 7(b)中目标检测网络的效果，本文方法将关注点集中在图中主要的物体上，对目标检测网络提取的特征中包含的冗余信息有明显的抑制效果。

(a) 原图 (b) ObjectNet (c) 本文方法

图 7 冗余信息的抑制效果

Fig. 7 Suppression effect of redundant information

2.3 消融实验

CCM 模块和 CG 模块可分别作为网络的一部分来改善网络性能。为了测试本文方法分别在 CCM、CG 以及它们的组合情况下的性能，本文通过它们的多种组合来分析模型在 MIT Indoor67 和 SUN397 两个数据集上的识别准确率，如表 5 所示。

根据表 5 的数据显示，在 MIT Indoor67 数据集上，网络在分别只有 CCM 和 CG 的情况下，准确率为 88.35% 和 87.91%，相较于二者都没有的情况下的 86.40%，识别准确率分别提高了 1.95 个百分点和 1.51 个百分点；而在同时具备 CCM 和 CG 时，即本文方法识别准确率达到了 90.28%。在 SUN397 数据集上，本文方法也取得了不错的效果，在只采用 CCM 或 CG 情况下，准确率分别提高了 1.55 个百分点和 2.39 个百分点；而

www.joca.cn

同时采用CCM和CG时,最终识别准确率达到了81.15%。以上实验表明,CCM和CG模块对室内场景识别产生了积极的影响,提高了识别的准确率。

表5　CCM和CG的不同组合效果对比

Tab. 5　Effect comparison of different combination of CCM and CG

目标特征之后	CCM之后	识别准确率/%	
		MIT Indoor67	SUN397
—	—	86.40	75.81
CCM	—	88.35	77.36
—	CG	87.91	78.20
CCM	CG	90.28	81.15

3　结语

本文针对室内场景识别中目标特征与场景特征性质和维度不一致、特征信息冗余等问题,在结合目标检测和场景识别的基础上,提出一种改进的室内场景识别方法。通过CCM将场景图像中的目标特征转换为场景特征,使得二者具有相同的特征维度,减少特征融合时的特征信息丢失;然后,利用CG抑制特征中的冗余信息,降低了不相关信息对场景识别的影响,提高了目标特征在室内场景识别中的作用。最后,本文分别在Scene-15、MIT Indoor67和SUN397三个场景数据集上进行了实验,并通过和其他方法对比,验证了本文方法有效提高了室内场景识别的准确率。

参考文献(References)

[1] LÓPEZ-CIFUENTES A, ESCUDERO-VIÑOLO M, BESCÓS J, et al. Semantic-aware scene recognition[J]. Pattern Recognition, 2020, 102: No. 107256.

[2] SONG X H, JIANG SH Q, HERRANZ L, et al. Learning effective RGB-D representations for scene recognition [J]. IEEE Transactions on Image Processing, 2019, 28(2): 980-993.

[3] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C]// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2006: 2169-2178.

[4] FEICHTENHOFER C, PINZ A, WILDES R P. Temporal residual networks for dynamic scene recognition [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7435-7444.

[5] 田艳玲,张维桐,张锲石,等. 图像场景分类技术综述[J]. 电子学报, 2019, 47(4): 915-926. (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.)

[6] ZHAO Z Y, LARSON M. From volcano to toyshop: adaptive discriminative region discovery for scene recognition [C]// Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 1760-1768.

[7] 储珺,苏亚伟,王璐. 自适应调节学习率和样本训练方式的场景分类[J]. 模式识别与人工智能, 2018, 31(7): 625-633. (CHU J, SU Y W, WANG L. Scene classification with adaptive learning rate and sample training mode [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(7): 625-633.)

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

[9] CHENG X J, LU J W, FENG J J, et al. Scene recognition with objectness[J]. Pattern Recognition, 2018, 74: 474-487.

[10] HERRANZ L, JIANG S Q, LI X Y, et al. Scene recognition with CNNs: objects, scales and dataset bias [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 571-579.

[11] WANG L M, GUO S, HUANG W L, et al. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs [J]. IEEE Transactions on Image Processing, 2017, 26(4): 2055-2068.

[12] REZANEJAD M, DOWNS G, WILDER J, et al. Scene categorization from contours: medial axis based salience measures [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4111-4119.

[13] SUN N, LI W L, LIU J X, et al. Fusing object semantics and deep appearance features for scene recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(6): 1715-1728.

[14] DU D P, WANG L M, WANG H L, et al. Translate-to-recognize networks for RGB-D scene recognition [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 11828-11837.

[15] 边小勇,费雄君,穆楠. 基于尺度注意力网络的遥感图像场景分类[J]. 计算机应用, 2020, 40(3): 872-877. (BIAN X Y, FEI X J, MU N. Remote sensing image scene classification based on scale-attention network [J]. Journal of Computer Applications, 2020, 40(3): 872-877.)

[16] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464.

[17] KHAN N, CHAUDHURI U, BANERJEE B, et al. Graph convolutional network for multi-label VHR remote sensing scene recognition[J]. Neurocomputing, 2019, 357: 36-46.

[18] KIM H J, FRAHM J-M. Hierarchy of alternating specialists for scene recognition [C]// Proceedings of the 2018 European Conference on Computer Vision, LNCS 11215. Cham: Springer, 2018: 471-488.

[19] WANG Z, WANG L M, WANG Y L, et al. Weakly supervised PatchNets: describing and aggregating local patches for scene recognition[J]. IEEE Transactions on Image Processing, 2017, 26(4): 2028-2041.

[20] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

This work is partially supported by the Surface Program of National Natural Science Foundation of China (61876121), the Natural Science Research Program of Jiangsu Province Higher Educations (19KJB520054).

XU Jianglang, born in 1996, M. S. candidate. His research interests include image processing, deep learning, scene recognition.

LI Linyan, born in 1983, M. S., senior engineer. Her research interests include multimodal information processing.

WAN Xinjun, born in 1996, M. S. candidate. His research interests include image processing, deep learning.

HU Fuyuan, born in 1978, Ph. D., professor. His research interests include image processing, pattern recognition, information security.

www.joca.cn