文章编号:1671-6833(2021)01-0042-08

结合特权信息与注意力机制的场景识别

孙 宁[1], 王龙玉[1,2], 刘佶鑫[1], 韩 光[1]

(1.南京邮电大学 宽带无线通信技术教育部工程研究中心,江苏 南京 210003;2.南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:在场景识别中,为了在只有 RGB 图像的测试阶段也能利用深度图像与 RGB 图像所包含的互补信息,以深度图像为特权信息,提出了一种端到端可训练的深度神经网络模型,用以结合特权信息和注意力机制。在该模型中,以图像编码到特征解码再到图像编码为架构,建立了由 RGB 图像到深度图像再到深度图像高层语义特征的映射关系。通过注意力机制,将 RGB 图像高层语义特征与对应的深度图像高层语义特征进行融合,输入分类网络,最终得到预测结果。在测试时,只需要输入 RGB 图像,便可在该模型获取的深度图像特权信息的帮助下,提升场景识别的性能。大量实验结果表明:本文方法在 SUN RGB-D 和 NYUD2 两个场景识别数据库中分别取得了 51.5%和 65.4%的识别正确率,验证了所提方法的有效性。

关键词:场景识别;特权信息;注意力;卷积神经网络

中图分类号:TN911.73 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2021.01.007

0 引言

场景图像由于其类内差异大和类间差异小的特点,一直是图像识别中一项颇具挑战性的任务。随着深度学习方法,特别是卷积神经网络(convolutional neural networks, CNN)的蓬勃发展,图像识别的性能在大规模图像数据库(如 ImageNet[1])的支持下已取得了显著的提高。为了支撑基于大规模深度神经网络的场景识别研究,各国研究人员也构建了一些大型的场景图像数据库。如 Places 数据库[2]和 Places2 数据库[3]等。其中 Places2 搜集了 365 种场景类别,图像数量超过 600 万张,是目前最大型的场景图像数据库。在这些大型数据库的帮助下,基于深度神经网络模型的场景识别方法[3]将场景识别的正确率推向了一个新的高度。

虽然大规模 CNN 模型辅以海量训练图像可以有效地提高场景识别的性能,但场景图像目标多、空间布局复杂、类间差异小的特点,使得完全依靠 RGB 图像一种模态的数据获得的场景识别性能与人类的认知能力仍有巨大的差异。

近年来,随着深度传感器的快速发展,将 RGB 图像与深度图像相结合的场景识别方法引起了研究人员的关注。研究结果表明,RGB 图像与深度图像之间具有明确且强烈的互补性。基于 RGB-D 双模态图像进行场景识别比使用单模态数据的方法有明显的优势[4-5]。

然而,深度图像获取不易:一方面,现有的 RGB-D 图像数据库相比 RGB 图像数据库要小得多;另一方面,在场景识别的实际应用中,输入算法的往往只有 RGB 图像。如果能将训练时由双模态数据学习得到的互补信息应用到只有单模态数据的测试阶段,将明显改善目前基于 RGB-D 图像的场景识别研究严重受到有限训练数据库制约的现状。

针对上述问题,本文将深度图像作为特权信息进行使用。所谓特权信息是指在训练时可用,测试时不可用的信息[6]。本文结合特权信息和注意力机制构建了一种端到端可训练的深度神经网络模型,称为 PIA-SRN (scene recognition network using privilege information and attention mechanism)模型。该模型通过图像转换(image-

收稿日期:2020-08-14;**修订日期**:2020-10-30

基金项目:国家自然科学基金资助项目(61471206, 61871445);江苏省优秀青年基金项目(BK20180088)

作者简介:孙宁(1978—),男,浙江温州人,南京邮电大学副教授,博士,主要从事图像处理、机器学习研究,E-mail:sunning@njupt.edu.cn。

to-image translation)建立 RGB 图像和深度图像的关系。该过程可以描述为 RGB 图像编码得到 RGB 图像高层语义特征,再进行特征解码得到深度图像,最后进行深度图像编码得到深度图像高层语义特征。为了进一步提升双模态数据互补信息的结合效果,使用注意力机制对双模态高层语义特征进行融合,最后得到场景的预测信息。PIA-SRN 在训练时基于双模态图像数据,在测试时只用输入 RGB 图像便可以取得更优的场景识别结果。

1　相关工作

1.1　基于 RGB-D 图像的场景识别

近年来,研究人员通过双流神经网络模型从 RGB 和深度图像中分别学习特定模态的特征,将两者融合后得出场景识别结果。Wang 等[5]以模块感知融合的方式从不同的模式中提取融合深层特征。Xiong 等[7]提出模态分离网络来提取模态一致性和特异性特征。然而,这些方法只考虑将颜色线索转移到深度网络,而忽略了深度线索也可以使 RGB 网络受益。Du 等[8]据此提出了 TRecgNet 模型,其在编码网络上连接一个解码模块,并用语义损失对其进行优化,实现了缺失模态的生成,同时优化了编码网络,提升了识别效果。TRecgNet 模型的核心是通过两个模态图像相互生成的过程,使各自的编码器网络学习有益的互补信息。从数据使用的角度来看,上述方法在训练和测试阶段都利用了 RGB-D 双模态图像进行场景识别。而本文方法则致力于结合特权信息和注意力机制实现训练时使用双模态数据,测试时只使用单模态数据的场景识别。

1.2　特权信息

特权信息(privilege information, PI)[5]是一种仅在训练期间可用而测试时无法获取的信息,已被应用于分类[5]、回归[9]等任务中。在计算机视觉中,许多信息可以被视为特权信息,例如文本和属性[9]等,这些信息根据具体问题设计和预定义。然而,现有的 PI 研究大多是基于支持向量机[5]的,随着深度学习研究的深入,研究人员开始探索基于 CNN 的特权信息使用方式。Hoffman 等[4]以深度图片作为特权信息,用幻觉网络的方式学习一种新的和互补的 RGB 图像表示。Garcia 等[10]同样使用深度图像作为特权信息对基于 RGB 图像的行为识别进行补充。在训练时,通过对抗学习与幻觉网络依据深度模态从 RGB 图像中模拟深度图像特征,从而在测试时缺失深度模态的状态下,提高 RGB 图像识别性能。本文利用双模态图像转换这种显性的操作,来隐性地实现双模态图像对单模态图像特征学习的有益补充,在训练阶段实现 PI 的嵌入,并在测试阶段发挥提升识别效果的作用。

1.3　注意力机制

图像识别算法中注意力机制的应用来自于对人类视觉系统的显著性研究。最近的工作证明,加入注意力机制的深度神经网络能提升多数图像识别任务的性能[11-12]。与针对单帧图像空间注意力建模和多帧图像时间注意力建模不同,本文方法中的注意力机制主要面向特征提取中的通道注意力。相关工作中,Wang 等[11]提出编码器解码器样式的残差注意力网络,通过在普通 Resnet[13]中增加由一系列卷积和池化操作组成的侧分支,突出特征图中有用特征,抑制无意义信息。Hu 等[12]提出加入门控机制的 SE-Net,挖掘通道间关系,选择性放大有价值的特征通道,抑制无用通道。本文考虑提取的特权信息特征与 RGB 特征间仍然存在着必然的联系,提出将注意力与特权信息相结合,促进彼此特征的提取,提高解码器生成特权信息的语义质量。

2　结合特权信息与注意力机制的场景识别网络

2.1　PIA-SRN 模型概述

本文以深度图像为特权信息,构建了一个端到端可训练的深度神经网络 PIA-SRN 来实现场景识别任务。在训练时,通过上述网络将深度图像的语义信息迁移到对 RGB 图像的特征学习中,提升网络对场景图像的特征提取。在测试时,训练后的上述网络在只有 RGB 图像的条件下,可以获得较没有特权信息嵌入的网络更优的识别效果。

PIA-SRN 模型主要由编码器网络(E-Net)、解码器网络(D-Net)、语义一致性网络(S-Net)、特权信息提取网络(PI-Net)、分类器网络(C-Net)和注意力模块(A-Mod)6 部分组成,如图 1 所示。其中语义一致性网络在训练时,通过语义一致性损失拉近生成深度模态图像与真实深度模态的语义,保证了编解码器网络生成图像的语义质量;在测试时编解码器网络将无须语义一致性网络的指导,为减少运算,语义一致性网络将会去除。PIA-SRN 模型的工作原理如图 1 所示。

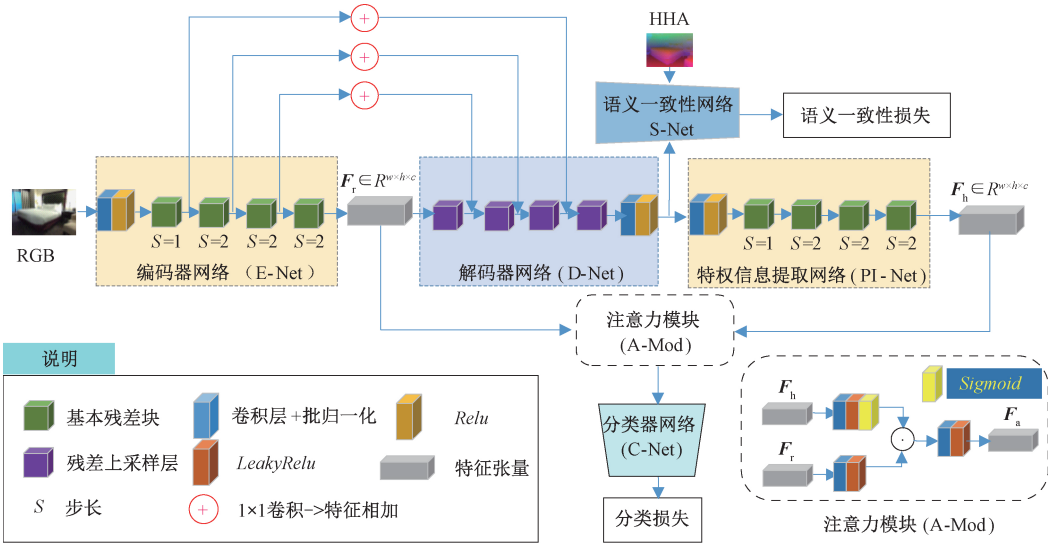

图1 PIA-SRN模型结构示意图

Figure 1 Schematic diagram of PIA-SRN model structure

从信息流程上看,RGB图像经过E-Net和D-Net后可以生成相对应的深度图像 $\boldsymbol{I}_{gh}$ 和由RGB图像学习得到的高层场景语义特征 $\boldsymbol{F}_r$。$\boldsymbol{I}_{gh}$ 输入PIA-SRN后,可以得到与 $\boldsymbol{F}_r$ 对应的由深度图像得到的高层场景语义特征 $\boldsymbol{F}_h$。$\boldsymbol{F}_r$ 和 $\boldsymbol{F}_h$ 经过A-Mod的处理后得到注意力加权后的高层场景语义特征 $\boldsymbol{F}_a$,输入C-Net后,便可以得到场景图像的识别结果。

对PIA-SRN模型的训练是一种多任务学习。一个学习目标是使得由 $\boldsymbol{I}_r$ 生成的 $\boldsymbol{I}_{gh}$ 与对应的 $\boldsymbol{I}_h$ 语义相一致;另一个学习目标则是由 $\boldsymbol{F}_a$ 经过C-Net之后输出的预测与RGB图像的场景类别相一致。在测试时,将新的RGB场景图像 $\bar{\boldsymbol{I}}_r$ 输入PIA-SRN,经过E-Net可得 $\bar{\boldsymbol{F}}_r$,再经过D-Net和PI-Net后得到合成的 $\bar{\boldsymbol{F}}_h$,再经过A-Mod和C-Net便能得到场景图像识别的结果。由此,本文方法将深度图像视为特权信息,为测试时仅使用RGB图像的场景识别提供了有益的信息,提升了单模态图像场景识别的正确率。

2.2 PIA-SRN模型的组成部分

2.2.1 编码器网络(E-Net)

编码器网络E-Net,由预训练好的Resnet18模型中的特征提取部分所构建,图1中 S 表示Resnet基本残差块中的步长。编码器输入尺寸为224×224×3的RGB图像 $\boldsymbol{I}_r$,输出14×14×512特征图 $\boldsymbol{F}_r$。为了防止信息的丢失,去除了Resnet18中的池化层。经过卷积核尺寸为7,步长为2的卷积块后通过4个基本残差块输出 $\boldsymbol{F}_r$,作为解码器网络的输入。

2.2.2 解码器网络(D-Net)

解码器网络分别由残差上采样模块和7×7卷积模块组成。输入为 $\boldsymbol{F}_r$,输出为224×224×3的深度图像 $\boldsymbol{I}_{gh}$。为了提高模型在训练时的有效性,编码器与解码器间通过1×1卷积跳跃连接(skip connection)。编码器输出 $\boldsymbol{F}_r$ 作为解码器的输入,每经过一次残差上采样,尺寸扩大2倍,最终生成深度图像 $\boldsymbol{I}_{gh}$,作为特权信息提取网络输入,在训练阶段也作为语义一致性网络的输入。

2.2.3 语义一致性网络(S-Net)

语义一致性网络采用了文献[8]中的网络结构,训练时解码器由 $\boldsymbol{F}_r$ 生成 $\boldsymbol{I}_{gh}$ 提供损失信息。其同样使用Resnet18模型为基础,语义一致性网络输入为深度图像 $\boldsymbol{I}_{gh}$ 与真实深度模态 $\boldsymbol{I}_g$,输出语义一致性损失为 $L_{content}$。为更有效地拟合深度图像特征,本文先使用深度图片对Resnet18网络进行了预训练。本文摒弃了使用 $L1$ 损失来生成深度图像 $\boldsymbol{I}_{gh}$ 与真实深度模态 $\boldsymbol{I}_g$ 之间的语义,使用幻觉损失[4](hallucination loss)的方法来拉近每层特征。本文设定生成图像为 $\boldsymbol{I}_{gh}$,特权信息为 $\boldsymbol{I}_h$,语义一致性损失为:

$$L_{content}(\boldsymbol{I}_{gh},\boldsymbol{I}_g,j)=\frac{1}{L}\sum_{i=1}^{L}\| \sigma(\boldsymbol{\Phi}^j_{I_{gh}})-\sigma(\boldsymbol{\Phi}^j_{I_g})\|_2^2。\quad(1)$$

式中:σ 为 *Sigmoid* 函数,$\sigma = 1/(1+e^{-x})$;$\boldsymbol{\Phi}^j_{I_{gh}}$、$\boldsymbol{\Phi}^j_{I_g}$ 分别为数据 $\boldsymbol{I}_{gh}$、$\boldsymbol{I}_g$ 在Resnet 18中的第 j 层基本残差块的输出特征;L 为基本残差块数量。

2.2.4 特权信息提取网络(PI-Net)

特权信息提取网络与编码器网络结构一致,但作用的对象不同:编码器网络从RGB图像中提取高层语义特征 F_r,特权信息提取网络则是从深度图像中提取高级语义特征 $\boldsymbol{F}_h$,其输入为 $\boldsymbol{I}_{gh}$,输出为 $\boldsymbol{F}_r$。

2.2.5　注意力模块(A-Mod)

注意力模块实现特权信息与单模态特征的结合,用于获取一组语义加权特征 $\boldsymbol{F}_a$,然后输出给分类器网络,得到最终场景分类结果,结构如图 2 所示。本文中采用了门控点乘的方式结合 $\boldsymbol{F}_h$ 与 $\boldsymbol{F}_r$。为了让 Attention 层适应特权特征和 RGB 特征,在 Attention 层前二者分别进行了一次卷积。$\boldsymbol{F}_h$ 卷积后通过 *Sigmoid* 激活层生成了权重矩阵 $\boldsymbol{F}_{h1}$,选用 *Sigmoid* 的目的是将结果限制在 0~1,以便于实现对 $\boldsymbol{F}_{r1}$ 的结果进行加权。$\boldsymbol{F}_{h1}$ 如式(2)所示:

$$\boldsymbol{F}_{h1} = \sigma(\varphi(w_h^{A_1}\boldsymbol{F}_h + b_h^{A_1})), \tag{2}$$

$\boldsymbol{F}_r$ 卷积后获得特征 $\boldsymbol{F}_{r1}$:

$$\boldsymbol{F}_{r1} = \varphi(w_r^{A_1}\boldsymbol{F}_r + b_r^{A_1})。 \tag{3}$$

式中:σ 和 φ 分别为 *Sigmoid* 和 *LeakyRelu* 激活函数;$w_h^{A_1}$、$w_r^{A_1}$、$b_h^{A_1}$、$b_r^{A_1}$ 分别为卷积层权重与偏置。$\boldsymbol{F}_{h1}$ 与 $\boldsymbol{F}_{r1}$ 点乘,消除了 $\boldsymbol{F}_{r1}$ 中的不相关信息,强调了有用信息。同理,在输入分类网络前也进行了一次卷积,输出 $\boldsymbol{F}_a$ 以适应分类器网络。

$$\boldsymbol{F}_a = \varphi(w_A^{A_2}(\boldsymbol{F}_{r1} \odot \boldsymbol{F}_{h1}) + b_A^{A_2})。 \tag{4}$$

式中:$w_A^{A_2}$、$b_A^{A_2}$ 为卷积层权重与偏置。

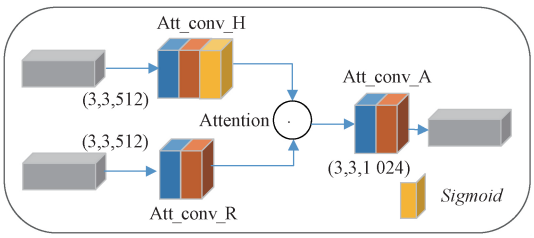

图 2　注意力模块结构示意图

Figure 2　Schematic diagram of attention module

2.2.6　分类器网络(C-Net)

分类器网络实现对 $\boldsymbol{F}_a$ 的最终特征学习和类别预测,由全局平均池化和 3 个全连接层构成。输入为 $\boldsymbol{F}_a$,输出为分类结果。全局平均池化维度为 14×14×1 024,最后一层全连接层维度为类别数,经过 *softmax* 输出场景类别。

2.3　PIA-SRN 模型的多任务训练

本文包含两个任务:一是对 RGB 图像进行分类,二是要拉近生成深度图像 $\boldsymbol{I}_{gh}$ 与真实深度模态图像 $\boldsymbol{I}_g$ 之间的语义,利于特权信息提取网络提取高级语义特征。

对于网络参数的更新,通过多任务学习的方式进行学习,如式(5)所示:

$$L_{total} = \alpha L_{content} + \beta L_{cls}。 \tag{5}$$

式中:$L_{content}$ 为语义一致性损失;L_{cls} 为分类损失;α、β 为损失权重,本文中分类损失为交叉熵损失,经过多次实验,为突出语义一致的重要性,将 α 设为 15,β 设为 1。

另外,由于所使用数据库中存在数据不均衡现象,本文对交叉熵损失采用权重重分配的方式平衡类别分类效果,如式(6)所示:

$$L_{cls\_balacned} = \frac{1}{N}\sum_i -w(y_i)\log\frac{f(x_i)y_i}{\sum_j f(x_i)_j}; \tag{6}$$

$$w(y) = \frac{N(y) - N(c\_min) + \delta}{N(c\_max) - N(c\_min)}。 \tag{7}$$

式中:$N(y)$ 为类别 y 的图像数量;c_max 和 c_min 分别为训练数据类别中最多与最少的图像数量;δ 设置为 0.01。

3　实验与分析

在本节中,基于两个 RGB-D 场景图像数据库对本文方法进行全面的测试,并对实验结果进行了详细的分析。

3.1　实施细节

实验中,对深度图片采用 HHA 编码[14](水平视差、离地高度和重力角度)几何编码,以更好地表征深度图像信息。配对后的 RGB 图像与深度图像的尺寸先调整到 256×256。为缓解训练中出现的过拟合现象,进行了数据扩增操作,将图像随机裁剪为 224×224,并进行概率为 0.5 的随机水平翻转。使用 Adam 算法[15]作为优化器进行网络优化。设定训练样本的批尺寸(batch size)为 16;学习周期为 200 次迭代;初始学习率为 0.000 2,前 30 次迭代保持不变,之后进行线性衰减。实验结果中的正确率为 3 次实验的平均值。

本文方法使用 PyTorch 深度学习框架实现。所有实验在配置为 2.4 GHz,8 核英特尔至强处理器,128 GB 内存,2 块显存为 12 GB 的英伟达 Titan Xp 显卡的图像工作站中运行。

3.2　数据库介绍

SUN RGB-D 数据库是目前最大的 RGB-D 场景图像数据库。包含 10 355 张 RGB-D 图像对。这些图像来自不同的采集设备,如 Asus Xtion、Microsoft Kinect v1 和 Intel RealSense。本文遵循文献[16]中的标准实验设定,选取 19 个超过 80 张图像的主要场景类别进行实验。其中训练集共有 4 845 张,测试集共有 4 659 张。

NYU Depth v2 (NYUD2)数据库相对较小,包括 27 类室内类别。按照文献[17]中的标准划分,这些类别被分为 10 个类别,包括 9 个最常见的类别和 1 个代表其余类别的其他类别。按照标准划分方法,训练集、测试集分别划分为 795、654

张图像。

3.3 基于 SUN RGB-D 数据库的实验

3.3.1 特权信息有效性实验

这一小节测试了本文方法中特权信息的有效性。实验中使用了 4 种 RGB 高层特征与特权信息相融合的方法,分别为点乘、拼接、求和与基于门控点乘的注意力机制。比较了使用注意力前后网络的识别效果。通过表 1 的实验结果可以发现,本文将特权信息提取网络的特征与编码器网络的特征通过点乘、拼接、求和 3 种方式融合时,识别正确率分别比单独使用 Resnet18 取得的 47.4%要高 1.2、2.6 和 3.9 个百分点;如果采用门控点乘的注意力机制,则识别正确率可以达到 51.5%。由此,证明了特权信息和注意力机制对单模态图像识别性能的促进作用,验证了 PIA-SRN 模型可有效地将训练阶段学习的特权信息融合进单模态特征中。

表 1 SUN RGB-D 数据库中的特权信息有效性实验结果

Table 1 Experimental results of privilege information validity in SUN RGB-D dataset

方法	使用注意力	PI 结合方式	正确率/%
Resnet18	否	否	47.4
PIA-SRN	否	点乘	48.6
	否	拼接	50.0
	否	求和	51.3
	是	门控点乘	51.5

3.3.2 与基于 RGB-D 双模态数据的场景识别方法的对比

目前几乎没有公开报道的场景识别方法如本文方法那样在训练阶段以深度图像为特权信息,一般在测试阶段只使用 RGB 图像。只有在文献[8]中的 TRecgNet 方法有在 SUN RGB-D 数据库上只使用 RGB 图像测试的结果。其他绝大多数的方法在训练和测试时都使用了 RGB-D 双模态数据。尽管如此,将本文算法的结果与基于 RGB-D 双模态数据的场景识别方法进行比较,由此进一步证明特权信息的使用对于单模态图像场景识别具有促进作用。

根据测试时使用的数据不同,表 2 中列出了本文方法与 3 种只使用 RGB 图像的方法,以及 10 种使用了 RGB-D 图像的方法的比较结果。可以看出,在只使用 RGB 图像进行测试的方法中,本文方法取得了最优的识别正确率,并明显地高于其他 3 种方法。与使用 RGB-D 图像进行测试的方法相比,本文方法的正确率比其中的 4 种方法的要高,与 2 种方法的正确率很接近,低于另外 4 种方法的结果。一方面,再次证明了本文方法在训练阶段学习到的 RGB-D 的互补信息,在测试时发挥了明显的作用,在只使用 RGB 图像进行识别时,接近甚至超过了一半以上的使用 RGB-D 两种图像方法的识别性能;另一方面,本文方法只使用了全局图像特征,未采用局部特征等方式深入挖掘图像中上下文语义,或是使用 RGB-D 图像对进行模态间差异性的学习等技巧,正确率低于最新的几种 RGB-D 场景识别方法。这也是本文方法下一步的重点研究方向。

表 2 本文方法与基于双模态数据方法在 SUN RGB-D 数据库上的比较

Table 2 Comparison of the method in this paper with the two-modal methods in SUN RGB-D dataset

方法	测试时使用数据库	正确率/%
Resnet18	RGB	47.4
TRecgNet[8]	RGB	49.8
TRecgNet Aug[8]	RGB	50.6
本文方法	RGB	51.5
SSCNN[18]	RGB-D	41.3
文献[11]的方法	RGB-D	41.5
文献[6]的方法	RGB-D	48.1
双流 Resnet	RGB-D	50.8
文献[19]的方法	RGB-D	52.4
文献[20]的方法	RGB-D	52.3
文献[21]的方法	RGB-D	53.3
DF^2Net[22]	RGB-D	54.6
MSN[7]	RGB-D	56.2
TRecgNet Aug [8]	RGB-D	56.7

本文在图 3 中给出了在 SUN RGB-D 数据库中的混淆矩阵。通过观察混淆矩阵可以发现,本文采用了权重重分配等方法,已经大大改善了由于 SUN RGB-D 中类别不均衡较为严重,在场景分类时易误分类比例较高的现象,如 bedroom。但由于有的类别如 study_space 类别数量确实过少,一定程度上还是有误分类的现象,这也将是本文接下来的工作重点。

3.4 基于 NYUD2 数据库的实验

3.4.1 特权信息有效性实验

与 SUN RGB-D 类似,本文在 NYUD2 数据库上也进行了特权信息有效性的实验。结果如表 3 所示。可以看出,在 NYUD2 库中所得的结果与在 SUN RGB-D 库中结果一致,再次验证了特权信息的有效性。同时值得注意的是,两次实验求和融合的结果都比较接近门控点乘的注意力融合的结果。这是由于 E-Net 和 D-Net 间跳线连接

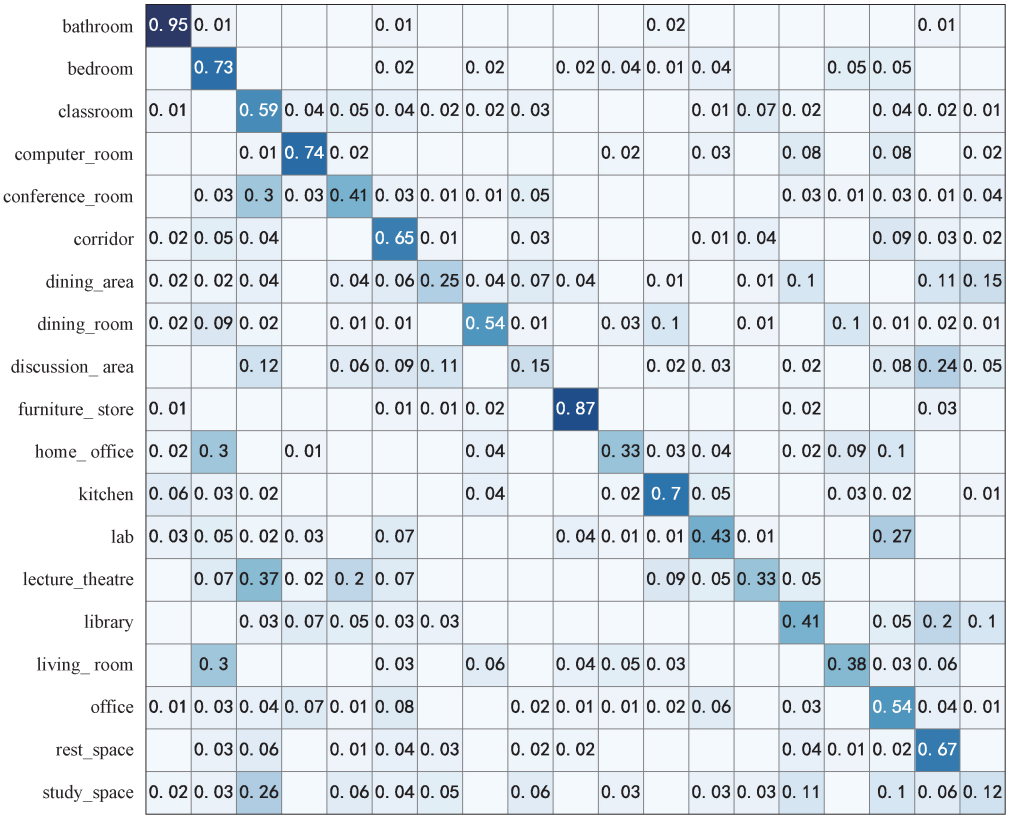

图 3　本文方法在 SUN RGB-D 中的混淆矩阵

Figure 3　The confusion matrix of this method in SUN RGB-D

时采用的是相加的方式,特权信息特征更容易与 RGB 图像特征相匹配。

表 3　NYUD2 数据库中的特权信息有效性实验结果

Table 3　Experimental results of privilege information validity in NYUD2 dataset

方法	PI 结合方式	正确率/%
Resnet18	否	59.8
PIA-SRN	点乘	60.2
	拼接	60.6
	求和	62.3
	门控点乘	62.8

3.4.2　基于 RGB-D 双模态数据的场景识别方法对比

同样,本文方法在 NYUD2 数据库上与其他 12 种方法的结果进行了比较,如表 4 所示。由结果可以看出,本文方法直接在 NYUD2 数据库上训练得到的结果要低于 TRecgNet Aug。这主要是由 NYUD2 数据库过小(训练图像只有 795 张)导致的过拟合现象。所以本文采用在 SUN RGB-D 数据库上进行预训练的策略来缓解过拟合问题,由此得到了最高为 65.4% 的识别正确率。与其他使用 RGB-D 图像的方法相比,本文方法要优于其中的 2 种,与其中 1 种正确率一致,低于另外4 种方法。与在 RGB-D 数据库上得到的结果趋势相同,也进一步验证了本文方法的有效性。同时,本文给出了在 NYUD2 数据库中进行场景分类的混淆矩阵,如图 4 所示,以便于观察实验结果。可以看出,NYUD2 中类别间相对均衡,相较于 SUN RGB-D 中的实验结果也有了进一步的改善。

表 4　本文方法与基于双模态数据方法在 NYUD2 数据库上的比较

Table 4　Comparison of the method in this paper with the two-modal method in NYUD2 dataset

方法	测试时使用数据	正确率/%
Resnet18	RGB	59.8
TRecgNet[8]	RGB	60.2
TRecgNet Aug[8]	RGB	64.8
TRecgNet*[8]	RGB	63.8
TRecgNet Aug*[8]	RGB	64.8
本文方法	RGB	62.8
本文方法*	RGB	65.4
双流 Resnet	RGB-D	63.8
文献[6]的方法	RGB-D	63.9
DF^{2}Net[22]	RGB-D	65.4
文献[19]的方法	RGB-D	65.8
文献[20]的方法	RGB-D	66.9
MSN[7]的方法	RGB-D	68.1
TRecgNet Aug[8]	RGB-D	69.2

注:* 表示使用 SUN RGB-D 进行预训练。

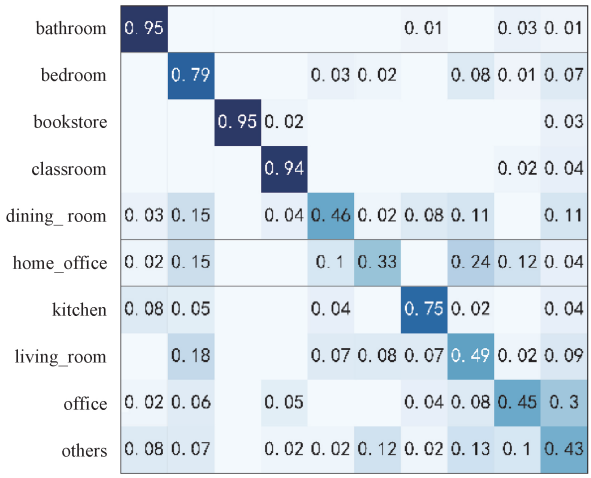

图4 本文方法在NYUD2中的混淆矩阵

Figure 4 The confusion matrix of this method in NYUD2

4 结论

提出了一种端到端可训练的深度神经网络模型(PIA-SRN)。该模型结合了特权信息和注意力机制,在训练阶段学习RGB-D双模态图像特征,在测试阶段仅使用RGB图像进行场景识别。一定程度上缓解了深度模态图片难以获取,RGB图像特征提取不充分的现状。深度图像以特权信息的方式,提升了单模态RGB图像进行场景识别时的正确率,达到了多数使用RGB-D双模态图像识别的效果。在两个公开的RGB-D场景识别数据库SUN RGB-D与NYUD2上验证了本文方法的有效性。

参考文献:

[1] JIA D, WEI D, RICHARD S, et al. Imagenet: a large-scale hierarchical image database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.

[2] ZHOU B L, LPEDRIZA A, XIAO J X, et al. Learning deep features for scene recognition using places database [J]. Advances in neural information processing systems, 2015,1: 487-495.

[3] ZHOU B, 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.

[4] HOFFMAN J, GUPTA S, DARRELL T. Learning with side information through modality hallucination[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 826-834.

[5] WANG A, CAI J, LU J, et al. Modality and component aware feature fusion for RGB-D scene classification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 5995-6004.

[6] VAPNIK V, VASHIST A. A new learning paradigm: learning using privileged information[J]. Neural networks, 2009, 22(5/6): 544-557.

[7] XIONG Z T, YUAN Y, WANG Q. MSN: modality separation networks for RGB-D scene recognition[J]. Neurocomputing, 2020, 373: 81-89.

[8] DU D, WANG L, WANG H, et al. Translate-to-recognize networks for RGB-D scene recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 11836-11845.

[9] SHARMANSKA V, QUADRIANTO N, LAMPERT C H. Learning to rank using privileged information[C]// 2013 IEEE International Conference on Computer Vision. New York: IEEE, 2013: 825-832.

[10] GARCIA N C, MORERIO P, MURINO V. Learning with privileged information via adversarial discriminative modality distillation[J]. IEEE transactions on pattern analysis and machine intelligence, 2020,42(10): 2581-2593.

[11] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 3156-3164.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.

[13] 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. New York: IEEE, 2016: 770-778.

[14] GUPTA S, GIRSHICK R, ARBELÁEZ P, et al. Learning rich features from RGB-D images for object detection and segmentation[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 345-360.

[15] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2020-05-15]. https://arxiv.org/abs/1412.6980.

[16] SONG S, LICHTENBERG S P, XIAO J. Sun RGB-D: a RGB-D scene understanding benchmark suite[C]// Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition. New York: IEEE, 2015: 567-576.

[17] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images [C]//European Conference on Computer Vision. Berlin: Springer, 2012: 746-760.

[18] LIAO Y, KODAGODA S, WANG Y, et al. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks[C]//2016 IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2016: 2318-2325.

[19] SONG X, HERANZ L, JIANG S Q. Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs[EB/OL]. (2018-01-21) [2020-05-15]. https://arxiv.org/abs/1801.06797.

[20] SONG X H, JIANG S Q, HERRANZ L. Combining models from multiple sources for RGB-D scene recognition [C]// International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI, 2017: 4523-4529.

[21] DU D, XU X, REN T, et al. Depth images could tell us more: enhancing depth discriminability for RGB-D scene recognition[C]//2018 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2018: 1-6.

[22] LI Y B, ZHANG J G, CHENG Y H, et al. DF^{2}Net: discriminative feature learning and fusion network for RGB-D indoor scene classification[C]//The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI). New Orleans: AAAI, 2018:7041-7048.

Scene Recognition Based on Privilege Information and Attention Mechanism

SUN Ning[1], WANG Longyu[1,2], LIU Jixin[1], HAN Guang[1]

(1.Engineering Research Center of Wideband Wireless Communication Technology of Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2.School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In the scene recognition, in order to use the complementary information contained in the depth images and the RGB images in the test phase with only RGB images, this paper used the depth image as the privilege information, and proposed an end-to-end trainable deep neural network model to combine the privilege information and attention mechanism. In the proposed method, the image encoding, feature decoding and then image encoding were used as the framework to establish a mapping relationship from RGB images to depth images and to high-level semantic features of depth images. By using of the attention mechanism, the high-level semantic features of RGB images were fused with the corresponding high-level semantic features of the depth image. And these two features were fed into the classification network to make the final prediction. In the test phase, only RGB images would be used, and the performance of scene recognition could be improved with the help of privilege information extracted from depth image. Through a large number of experiments, the method in this paper achieved 51.5% in the SUN RGB-D scene identification database and 65.4% in NYUD2 database, which verified the effectiveness of the method in this paper.

Key words: scene recognition; privilege information; attention mechanism; convolutional neural network