DOI: 10.13382/j.jemi.B2206084

融合 CNN 和 Transformer 的机器人室内场景识别*

刘 铁 段 勇
(沈阳工业大学信息科学与工程学院 沈阳 110870)

摘 要:为了提高机器人在复杂的室内环境中场景识别的准确率,本文提出一种融合卷积神经网络(convolutional neural network,CNN)和视觉 Transformer 结构的机器人室内场景识别模型。本文模型利用 CNN 提取场景局部特征,然后使用视觉 Transformer 结构捕捉特征中远距离依赖关系,其中提出的视觉 Transformer 结构包括 3 个部分,分别是特征编码结构(Attention Embedding)、Encoder 结构和一个将高层语义特征转化成像素级特征的结构(Attention Project)。本文研究的机器人场景识别模型利用 CNN 提高视觉 Transformer 局部细节特征的描述能力,同时通过视觉 Transformer 帮助 CNN 构建远距离特征的依赖关系,从而能够有效的表征和利用机器人工作场景图像的视觉特征。最后,通过机器人在实际工作环境中采集的数据集和开源的 COLD 数据集进行实验,验证了本文研究模型的有效性,场景识别精度更高。

关键词: CNN;Transformer;机器人;场景识别;局部特征
中图分类号: TP242;TN98 **文献标识码**: A **国家标准学科分类代码**: 520.2

Robot indoor scene recognition based on fusion of CNN and Transformer

Liu Tie Duan Yong
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: In order to improve the accuracy of robot scene recognition in complex indoor environments, this paper proposes a robot scene recognition model that fuses convolutional neural network (CNN) and visual Transformer structure. The model uses CNN to extract local features of the scene. And the visual Transformer structure is used to capture the distant dependencies in the features. The proposed visual Transformer structure consists of three parts, they are a feature encoding structure (Attention Embedding), an Encoder structure, and a structure that converts high-level semantic features into pixel-level features (Attention Project). The robot scene recognition model studied in this paper uses CNN to improve the description ability of local detail features of the visual Transformer. Furthermore, the visual Transformer helps CNN to construct the dependencies of distant features, which can effectively characterize and utilize the visual features of the robot working scene images. Finally, the effectiveness of the model is verified by experimenting with the dataset collected by the robot in the actual working environment and the open source COLD dataset. The scene recognition accuracy of our model is higher.

Keywords: CNN; Transformer; robot; scene recognition; local feature

0 引 言

随着机器人技术以及人工智能的飞速发展,机器人在机器学习、深度学习等技术的帮助下不断构建更为强大的功能体系以适用不同领域的应用场景。对于机器人而言,自主导航避障是能否正常完成任务的关键,而对周围环境场景的认知和理解程度直接影响机器人执行各项任务的准确性[1-2],因此,如何有效的提高机器人对周围环境的感知能力逐渐成为机器人研究领域的热点问题。

机器人场景识别是指机器人通过视觉传感器采集周围环境信息,对周围环境的场景进行认知和理解的过程,

收稿日期: 2022-12-06 Received Date: 2022-12-06
*基金项目:辽宁省高等学校优秀科技人才支持计划(LR15045)、辽宁省教育厅科学研究经费面上项目(LJKZ0139)资助

机器人工作所在的室内场景因内容复杂、物体繁多、物体遮挡、光线不足、不同场景相似度大等问题,因此准确的识别出机器人当前所处的室内场景具有一定挑战性。目前场景识别方法主要分为传统方法和深度学习方法两大类。传统场景识别方法首先利用图像表观和空间信息提取特征,然后利用机器学习器进行判别,受室内场景复杂、差异不大等因素的影响,现有的方法存在准确率低、实时性差等问题。相比较传统的场景识别方法,深度学习的方法能够提取更深层次的场景特征,实现输入场景图像后直接端对端的输出识别结果。有些研究者使用卷积神经网络(convolutional neural network,CNN)提取场景图像的特征,并对提取的特征进行分类,例如张杰[3]基于ResNet18[4]网络构建场景分类的模型;王桥等[5]提出一种利用CNN提取场景图像多尺度特征,并使用迁移学习方式训练多尺度融合的深度学习模型,不仅降低了训练成本,而且提高了场景识别精度;王彬等[6]提出一种包含卷积和反卷积结构的场景识别网络,使用逐层卷积上采样和跳层反卷积相结合的方法,提升网络的识别精度。然而在这些传统的CNN方法中,卷积算子擅长处理图像的局部特征,对于图像中的全局特征需要更深的网络结构才能捕获,带来了巨大的计算代价。

近年来,受自然语言处理领域的Transformer[7]模型的启发,一些研究者将Transformer架构应用在视觉领域,并且也取得了一定的成功。例如Parmar等[8]提出Image Transformer模型用来解决视觉领域的图像生成任务,将Transformer架构迁移到视觉领域,并取得当时先进的水平。最近Dosovitskiy等[9]提出一个纯Transformer结构的ViT模型,将图像拆分成块状序列输入到模型中,同样取得较高的水平。视觉的Transformer模型除了在图像分类任务取得先进水平之外,还包括目标检测、图像分割和视频处理[10]等任务。例如Carion等[11]将Transformer架构应用在目标检测任务上,不同于传统的循环神经网络检测目标的方法,该模型在候选框提取过程中使用Transformer结构直接预测物体的类别和位置信息。在图像分割任务中,Zheng等[12]提出一种序列到序列Transformer结构的模型,提高了模型提取全局语义信息的能力。

视觉Transformer构成的ViT[9]模型,其中多头注意力机制可以充分获取远距离的依赖关系,但是对图像局部特征的提取能力不是很强,在没有大规模数据预训练的条件下,无法发挥出模型的真实性能。对此,Wu等[13]提出VT模型,首先使用CNN结构提取图像特征,然后使用Transformer结构提取深层语义特征;Dascoli等[14]提出利用门控位置自注意力(gated positional self-attention,GPSA)结构将CNN本身的归纳偏置信息加入到Transformer结构,进而提高模型对样本的利用率,进而实现提高模型的识别精度。

针对机器人场景识别中传统的卷积结构不擅长捕获远距离依赖关系,导致机器人场景识别准确率低、模型复杂度大等问题,本文提出一种基于CNN和Transformer融合的机器人场景识别方法。CNN结构提取图像的浅层特征,补充Transformer中缺少的局部细节特征,Transformer结构捕捉远距离特征的依赖关系,帮助模型掌握图像的全局特征,CNN结构和Transformer结构之间互相补充结构短板,增强模型特征提取能力,提升模型在机器人场景识别中的准确率。

1　机器人室内场景数据的采集与处理

本文使用的是两轮驱动的先锋P3-DX移动机器人平台,如图1所示。其可以搭载视觉传感器、超声、激光等多种类型传感器。对机器人的控制是基于客户端软件高级机器人应用接口(advanced robot interface for application,ARIA)的基础上进行开发,在ARIA不仅提供了对机器人平台的简单控制,而且其内部封装了调用设备常用的类库。机器人在室内场景运动过程中,通过其搭载的视觉传感器采集周围环境的图像,然后通过通信串口将采集的RGB图像传输到客户端,在客户端进一步处理采集的数据,识别场景并输出结果。

图1　先锋P3-DX移动机器人平台

Fig. 1　Pioneer P3-DX mobile robot platform

P3-DX机器人搭载的是单目摄像头,在行进过程中不断采集周围环境图像,将全部图像构建成多段连续的图像序列。整个图像序列包括实验室、办公室、走廊等多个类别的场景图像,其中每一段图像序列中包括多帧连续的场景图像,使用关键帧算法提取出关键位置的图像,将提取的关键帧图像按照场景类别进行存储并构成机器人实际工作场景数据集。机器人采集的部分实际工作场景图像如图2所示。

服务端机器人的视觉传感器采集室内场景图像可能存在光线不足、物体遮挡、图像不清晰等问题,因此需要

图 2　部分机器人工作场景图像

Fig. 2　Some robot working scene images

对采集的图像进行初步的预处理操作。首先对图像进行直方图均衡化处理,均衡图像的对比度,其次对图像进行中值滤波处理,平滑图像中存在的噪声,然后按照场景类别中图像数量的比例进行筛选操作,最后将处理后的所有数据构建数据集,准备在模型训练中使用。

由于每一个场景中的图像相似度较大,为了提高模型的泛化性,对采集的图像进行数据增强。为了模拟真实环境中采集图像过程中出现的情况,本文随机对图像进行镜像翻转、左右翻转、旋转、缩放等仿射变换、高斯模糊、锐化、高斯噪声等操作,并将增强的图像添加到数据集中,扩充数据集的多样性。

2　模型结构

机器人室内场景识别是一个比较复杂的问题,因室内场景中出现的物体种类多、光线暗、环境复杂,因此不仅需要模型有较高的局部特征处理能力,而且需要模型能够构建特征之间依赖关系,具有较强的全局理解能力。

本文构建的模型包括 CNN 结构和 Transformer 结构,首先利用 CNN 提取机器人工作环境场景图像的局部特征,获取场景图像的浅层特征;然后使用 Transformer 结构学习浅层特征中远距离特征之间的依赖关系,获取图像中的深层语义信息,整体模型的框架如图 3 所示。

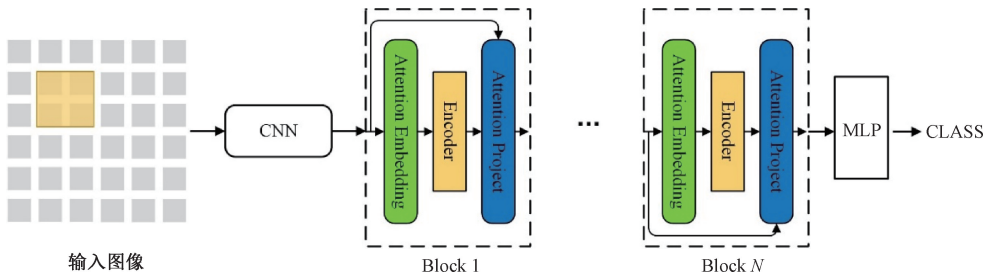


图 3　模型整体结构

Fig. 3　Overall structure of the model

如图 3 所示,本文模型先利用 CNN 结构提取机器人场景图像的局部特征,然后将局部特征输入到 Transformer 结构中。在 Transformer 结构中,首先使用 Attention Embedding 结构将场景图像转换成高层次语义特征,然后使用 Encoder 层构建特征之间的依赖关系,最后使用 Attention Project 结构将高层语义特征转化成像素级的语义特征。在 Transformer 结构后,会输出一个像素级别的特征,使用全连接层可以直接对输出的场景特征进行分类。

为了保障 Transformer 结构可以充分利用 CNN 提取的特征结构,在 Transformer 的 Encoder 层与 CNN 结构之间添加 Attention Embedding 结构和 Attention Project 结构,采用注意力机制的方法将视觉特征转换成 Transformer 可以进一步处理特征的结构。模型整体结构在 Transformer 模型和 VT[13] 模型的基础上进行优化,优化现有的 CNN 特征提取结构、Attention Embedding 结构和 Attention Project 结构。本文模型由一个 CNN 特征提取骨干网络和多个特征处理单元 Block 构成,CNN 模型的复杂度和 Block 的数量需要根据任务的复杂度动态调整。

2.1　CNN 结构

本文使用 EfficientNet[15] 作为特征提取骨干网络,EfficientNet 不仅在大型数据集上有较高的识别精度,模型提取特征能力更强,而且模型复杂度更低,参数量更少。并且其具有良好的缩放体系,根据任务的复杂程度不同,通过调整参数实现对网络的深度、宽度和图像的分辨率进行控制。EfficientNet 网络整体采用复合的缩放方法,通过复合系数 ϕ 同时缩放或者扩大深度、宽度和分辨率的缩放比例。

本文选择 EfficientNet_b1 作为机器人工作环境场景图像特征提取的基础网络,其中一共包括 16 个 MBConv 结构块,MBConv 是的结构如图 4 所示。其中包括卷积结构(convolutional layer, Conv)、标准化结构(batch normalization, BN)、激活函数(Swish)、深度可分离卷积结构(depthwise spearable convolution, DWConv)和 SENet[16] 注意力结构。

2.2　Attention Embedding 结构

CNN 模型擅长提取局部特征,但无法有效地获取远距离的依赖关系,而 Transformer 结构可以更好地处理远

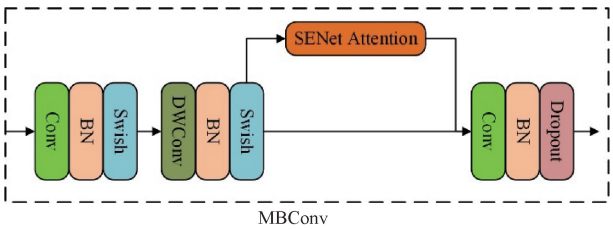


图4　MBConv结构

Fig. 4　Structure of the MBConv

距离的依赖关系，因此为了有效的利用CNN和Transformer的优点构建一个即能处理局部细节信息又能捕捉远距离依赖关系的模型，提出了Attention Embedding结构。该结构是CNN场景图像特征和Transformer模型之间的特征编码结构，将CNN提取的视觉特征转化为高层次的语义特征，然后Transformer可以充分利用这些语义特征捕捉语义中特征之间的依赖关系。本文在VT[13]模型的Recurrent Tokenizer结构上进行改进，优化了空间注意力的基础结构，具体结构如式(1)~(4)所示，其中式(1)为编码结构主要公式，式(2)~(3)分别是计算特征K、Q、V的公式。

$$T_n = Softmax\left(\frac{KQ^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{1}$$

$$K = X\boldsymbol{W}_k \tag{2}$$

$$Q = T_{n-1}\boldsymbol{W}_q \tag{3}$$

$$V = X\boldsymbol{W}_v \tag{4}$$

式中：$\boldsymbol{W}_k$、$\boldsymbol{W}_q$和$\boldsymbol{W}_v$都是需要学习的参数矩阵，d是隐藏层的维度，T_{n-1}是上一个Attention Embedding结构输出的特征，X是CNN模型输出的特征结构，K、Q、V这3个特征经过式(2)、(3)和(4)计算后代入到式(1)中计算结果。在空间注意力机制的帮助下，模型关注更重要的区域，而不是平等对待每一个区域，提高模型处理环境场景特征的效率。Attention Embedding结构是一个循环结构，当前层的输入同样需要上一层的输出。当前层的输入为上一个Block中Attention Project的输出，并且当前层还需要使用上一个Block中Attention Embedding输出的T_{n-1}来指导当前层T_n的计算，在此注意力计算的过程中使用上一层T_{n-1}指导当前层提取更深层的语义特征[13]。最后，将Attention Embedding处理后的特征输入Encoder层中，通过Transformer的注意力机制和编码机制捕捉当前特征中远距离依赖关系，获取对全局理解。

2.3　Attention Project结构

经过Transformer构建特征之间的依赖关系后，输出的一个具有视觉特征信息的高层次特征，然而视觉方面的任务中常需要像素级别的特征[13]，因此需要在模型中添加一个解码结构(Attention Project)，将高层次的语义特征转换成像素级别的视觉特征。通过一个跳层连接结构和注意力结构将带有视觉信息的高层次特征转换成具有像素级别的特征。Attention Project结构是一个类似解码的转化结构，在模型VT[13]的基础上优化注意力的结构，将Transformer捕获的依赖关系通过注意力机制计算后添加到原来CNN提取的特征中，突出特征中的重要区域，具体如式(5)所示。

$$Out = X + Softmax\left(\frac{(X\boldsymbol{W}_q)(T_n\boldsymbol{W}_k)^{\mathrm{T}}}{\sqrt{d}}\right)(T_n\boldsymbol{W}_v) \tag{5}$$

其中，$\boldsymbol{W}_k$、$\boldsymbol{W}_q$和$\boldsymbol{W}_v$都是需要学习的参数矩阵，d是隐藏层的维度，T_n是Attention Embedding结构最后输出的内容，X是CNN模型输出的特征结构。

3　实验结果及分析

本文模型主要在机器人场景识别领域进行应用，因此设计并实现一个解决该问题的准确率更高、时间复杂度低的深度学习模型。本文采集真实环境场景的机器人型号为先锋P3-DX机器人，使用的深度学习模型主要在深度学习服务器中进行训练，将训练好的深度学习模型应用于机器人进行场景识别。

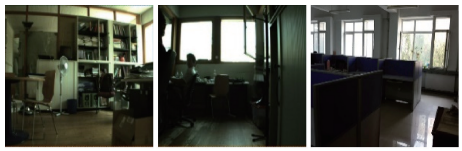
本文实验训练模型使用的优化器为Adam[17]，训练过程中输入模型的场景图像的大小为224×224，batch_size大小设置为16，Epoches大小设置为140，学习率设置为0.000 1，权重衰减为1×10^{-5}，模型中Transformer结构的Block数量设置2，Transformer中的heads数量设置为4，CNN特征提取模型中模型深度、宽度的缩放比例分别设置为1.1，1.2。

3.1　实验数据集

本文使用的实验数据集主要包括两个部分，其中第1部分是第1节中介绍的机器人采集的真实场景数据，例如实验室场景、走廊场景、办公室场景等，并将真实场景数据按照第1节介绍的预处理方法处理并构建数据集；第2部分是开源的COSy定位数据集(COSy localization database，COLD)数据集，该数据为Freiburg大学的自动智能系统实验室建立的机器人室内场景数据。数据集场景图像如图5和6所示，图5展示的是机器人采集后的场景图像，图6为对应的预处理后结果，其中图6(a)、(b)场景图像来自COLD数据集的场景，图6(c)场景图像来自本文机器人采集的真实场景图像。

为了提高模型的泛化能力，验证模型提取特征的能力，本文将两部分场景图像数据集构建成一个大的数据集，一共20类机器人场景，共5 439张图像，并将数据集按照6∶2∶2的比例划分成训练集、验证集和测试集。

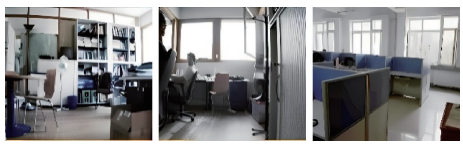
为了提高模型的泛化性，对数据集中的数据进行数据增强，将数据集分成两个版本，第1个版本保持原始数



(a) 数据集场景1 (b) 数据集场景2 (c) 真实场景
(a) Dataset scene 1 (b) Dataset scene 2 (c) Real scenes

图 5　原始场景图像

Fig. 5　Original scene image



(a) 数据集场景1 (b) 数据集场景2 (c) 真实场景
(a) Dataset scene 1 (b) Dataset scene 2 (c) Real scenes

图 6　预处理后场景图像

Fig. 6　Scene image after preproccessing

据集,不进行数据增强;第 2 个版本对数据进行数据增强,保存数据增强后的结果,保证每一个模型使用相同的数据集进行训练,并记录实验结果,分析数据增强对模型的影响。

3.2　实验分析

为了证明本文模型的有效性,在相同的数据集上对比了 ResNet50[4]、EfficientNet[15]、ViT[9]、VT[13] 等模型的准确率,实验结果如表 1 所示。其中准确率(Acc)指标表示模型的准确率,参数数量(Param)指标表示模型中参与训练的参数数量,浮点运算数(FLOPs)指标表示模型的复杂度,训练时间(Train_time)指标表示模型训练的时长,以 min 为单位。为了在实验中方便的表示模型,将本文模型表示为 EVT(efficientnet+vision transformer)。并且在实验中将 EfficientNet 结构替换成 SENet 结构,其模型表示为 SVT(SENet vision Transformer)。

表 1　不同模型的综合指标

Table 1　Comprehensive indicators of different models

模型	Acc/%	Param/M	FLOPs/G	Train_time/min
ResNet50[4]	81.64	25.50	6.57	10.0
EfficientNet_b1[15]	83.80	6.50	0.93	11.8
ViT[9]	79.20	85.00	16.00	25.5
VT[13]	84.10	12.11	1.45	18.0
SVT	85.90	12.81	1.46	21.4
EVT(本文)	87.59	4.60	0.81	20.2

从表 1 中可以看出本文的 EVT 模型有更高的准确率,并且该模型需要训练的参数数量和模型复杂度都比较低。表中的模型主要有 3 类,包括传统的 CNN 模型、视觉的 Transformer 结构模型和既有 CNN 又包括视觉 Transformer 的模型,从表 1 中可以看出,本文模型与传统的 CNN 模型相比,准确率更高,参数更少,说明通过 Transformer 结构可以更好地构建全局的依赖关系;与视觉 Transformer 的 ViT[9] 模型相比,在视觉的 Transformer 中加入 CNN 结构可以更好地帮助模型处理局部细节特征;与 VT[13] 模型相比,说明本文模型的改进更加有效,可以更好地利用局部构建对全局理解。对比模型 SVT 和本文 EVT 结构可知,CNN 的特征提取能力越强,提取的局部细节特征越丰富,在 Transformer 结构中捕捉的全局依赖关系更加准确,最后,模型的识别准确率更高。对比模型训练时间指标,可以发现带有 Transformer 结构的模型训练时间都比传统 CNN 模型训练时间长,但是本文 EVT 模型的训练时间在 Transformer 系列模型中趋于中等水平。

为了增加模型的泛化性,对数据集进行一定的数据增强处理,本实验使用的模型仍为 ResNet50[4]、EfficientNet[15]、ViT[9]、VT[13]、SVT 和本文的 EVT,数据增强前后的对比实验结果如图 7 所示,图中横坐标为不同模型,纵坐标为每个模型的准确率。

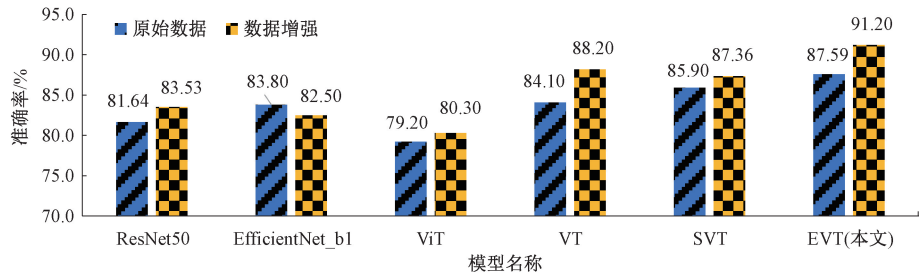


图 7　数据增强的不同模型的综合指标

Fig. 7　Comprehensive indicators of different models for data enhancement

从图 7 的实验结果可知,对于实验中使用的多数模型,数据增强对提高模型的泛化性有一定帮助,数据增强操作产生的样本增加数据集的多样性,增加数据之间的差异性,其中本文的 EVT 模型的准确率提升了 3.61%。但对于实验中的 EfficientNet_b1 模型,数据增强并没有提升模型的识别精度。

为了验证注意力机制在本文模型中的有效性,绘制了 VT[13]模型、EfficientNet[15]模型和本文 EVT 模型在训练过程中验证集的准确率曲线,具体如图 8 所示。从图中可以看出,视觉 Transformer 结构的模型在训练的初始化阶段模型的准确率超过了传统的 CNN 模型,说明注意力结构可以帮助模型关注场景图像中的重要环境信息。在训练中间部分,受模型参数数量和复杂度的影响,造成本文 EVT 模型收敛速度略慢,准确率略低于 VT[13]模型的准确率,但随着训练次数的增加,在收敛阶段本文模型准确率逐步的超越了 VT[13]模型的准确率。说明本文的 EVT 模型有效地将 CNN 和 Transformer 模型结构进行融合,使其能够更好地关注局部细节特征,以及更快速地构建远距离的依赖关系。改进的 Attention Embedding 和 Attention Project 更加有效的利用局部特征构建远距离的依赖关系,在场景识别问题中获得更高的准确率。

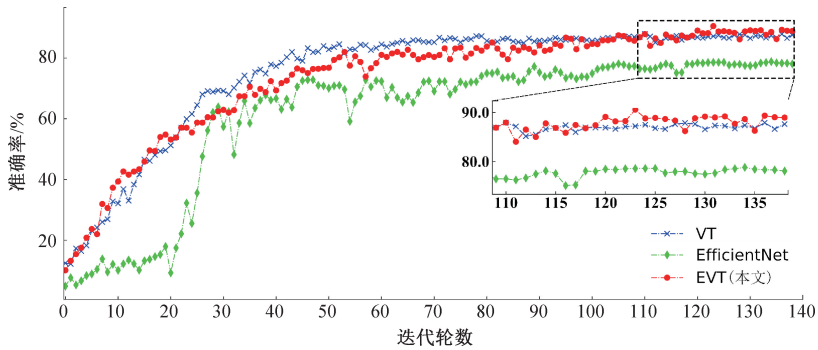
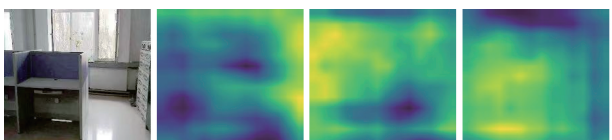


图 8　部分模型验证集准确率曲线

Fig. 8　Accuracy curve of partial model validation set

为了证明本文提出的模型可以充分利用 CNN 结构弥补 Transformer 结构忽略局部细节的缺点,将 CNN 结构的输出和 Transformer 结构的第 1 个 Block 输出结果进行可视化,如图 9 和 10 所示,分别给出 CNN 结构输出结果图和 Transformer 结构输出结果图。图 9 和 10 中(a)为输入模型的图像,(b)、(c)、(d)为模型中对应结构的输出结果,其中 Block 一共输出 512 个特征图,CNN 结构输出 256 个特征图,本文分别随机选取其中 3 个进行展示。

由图 9 可知,CNN 结构平等的对待所有特征,没有针对性的关注环境周围场景特征,而图 10 能够更好地反映出模型更加关注场景中的物体。通过对比可知,在注意力机制的帮助下,模型更加关注真实场景中周围环境物体,并且不同的特征图中注意力机制关注的重点不同。因此说明通过本文优化的模型可以充分利用 CNN 结构补充 Transformer 中缺失的细节信息,最终提升机器人场景识别准确性。

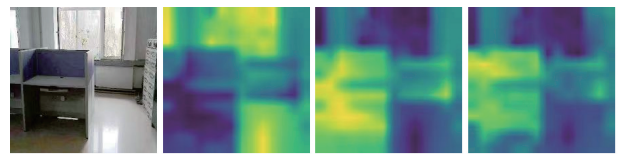


(a) 输入图像　(b) 输出图像1　(c) 输出图像2　(d) 输出图像3

(a) Input image　(b) Output image 1　(c) Output image 2　(d) Output image 3

图 9　模型 CNN 结构输出结果

Fig. 9　Model CNN structure output results



(a) 输入图像　(b) 输出图像1　(c) 输出图像2　(d) 输出图像3

(a) Input image　(b) Output image 1　(c) Output image 2　(d) Output image 3

图 10　模型 Block 1 结构输出结果

Fig. 10　Model block 1 structure output results

4　结　论

为了提高机器人在室内复杂环境中场景识别的准确率,本文提出一种新的 CNN 和 Transformer 融合的机器人场景识别模型。首先使用机器人采集室内环境信息,然后根据采集的场景信息构建训练模型的数据集,并且也添加 COLD 开源数据集,增加数据集的多样性,提高模型的泛化性。本文模型以 CNN 擅长处理局部特征和 Transformer 可以捕捉远距离依赖关系为思路,利用 CNN 弥补 Transformer 结构忽略局部细节的问题,利用 Transformer 结构帮助 CNN 构建远距离依赖关系。文本模型在 VT 模型的基础上进行优化,优化视觉特征提取结构,使用 EfficientNet 结构作为特征提取骨干网络;并且提出 Attention Embedding 和 Attention Project 结构,利用注意力机制将 CNN 提取的特征转化成高层次语义特征,使用 Transformer 结构构建依赖关系,然后使用 Attention Project 结构将高层次语义转化可以进行视觉分类任务像

素级特征,最后使用全连接层对特征进行分类。本文的模型在文中自建和开源的室内环境场景的数据上,识别精度高达 91.2%,并设置多组对比实验,证明优化的模型在解决机器人场景识别问题的有效性。

参考文献

[1] 刘明春. 基于深度学习的变电站巡检机器人道路场景识别[D]. 成都: 西南交通大学, 2019.
LIU M CH. Road scene recognition of substation inspection robot based on deep learning[D]. Chengdu: Southwest Jiaotong University, 2019.

[2] 张仪, 冯伟, 王卫军,等. 融合 LSTM 和 PPO 算法的移动机器人视觉导航[J]. 电子测量与仪器学报, 2022, 36(8): 132-140.
ZHANG Y, FENG W, WANG W J, et al. Visual navigation of mobile robots based on LSTM and PPO algorithms[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(8): 132-140.

[3] 张杰. 移动服务机器人室内场景识别关键技术研究[D]. 天津: 天津理工大学, 2022.
ZHANG J. Key technologies research on indoor scene recognition of mobile service robot[D]. Tianjin: Tianjin University of Technology, 2022.

[4] 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, 2016: 770-778.

[5] 王桥, 胡春燕, 李菲菲. 基于深度迁移学习与多尺度特征融合的场景识别方法[J/OL]. 电子科技: 1-9[2022-11-21].
WANG Q, HU CH Y, LI F F. Scene recognition algorithm based on deep transfer learning and multi-scale feature fusion[J/OL]. Electronic Science and Technology: 1-9[2022-11-21].

[6] 王彬. 深度学习机制下家庭服务机器人室内场景识别方法研究[D]. 南京: 南京邮电大学, 2019.
WANG B. Research on indoor scene recognition of home service robot based on deep learning[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2019.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[8] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]. International Conference on Machine Learning, 2018: 4055-4064.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[10] ZENG Y, FU J, CHAO H. Learning joint spatial-temporal transformations for video inpainting[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 6881-6890.

[11] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision, 2020: 213-229.

[12] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 6881-6890.

[13] WU B, XU C, DAI X, et al. Visual transformers: Token-based image representation and processing for computer vision[J]. arXiv preprint arXiv: 2006.03677, 2020.

[14] DASCOLI S, TOUVRON H, LEAVITT M L, et al. Convit: Improving vision transformers with soft convolutional inductive biases[C]. International Conference on Machine Learning, 2021: 2286-2296.

[15] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]. International Conference on Machine Learning, 2019: 6105-6114.

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

[17] KINGMA D, BA J. Adam: A method for stochastic optimization[C]. International Conference on Learning Representations, 2015: 1-15.

作者简介

刘铁,2020 年于沈阳工业大学获得学士学位,现为沈阳工业大学硕士研究生,主要研究方向为计算机视觉。
E-mail: ltsuperman163@163.com

Liu Tie received his B. Sc. degree from Shenyang University of Technology in 2020, M. Sc. candidate at Shenyang University of Technology now. His main research interest includes computer vision.

段勇(通信作者),沈阳工业大学信息科学与工程学院教授,博士生导师,主要研究方向为自主机器人、机器学习、计算机视觉。
E-mail: duanyong0607@126.com

Duan Yong (Corresponding author) is a professor and Ph. D. supervisor at School of Information Science and Engineering, Shenyang University of Technology. His main research interests include autonomous robot, machine learning and computer vision.