基于 scSE 非局部双流 ResNet 网络的行为识别

李占利 王佳莹 靳红梅 李洪安
(西安科技大学计算机科学与技术学院 陕西 西安 710600)

摘 要 针对双流网络对包含冗余信息的视频帧存在识别率低的问题,在双流网络的基础上引入 scSE(Spatial and Channel Squeeze & Excitation Block)和非局部操作,构建 SC_NLResNet 行为识别框架。该框架将视频划分为等分不重叠的时序段并在每段上稀疏采样,提取 RGB 帧以及光流图作为 scSE 模块的输入;将经过 scSE 处理的特征输入非局部双流 ResNet 网络中,融合各分段得到最终的预测结果。在 UCF101 以及 Hmdb51 数据集上实验准确率分别达到 96.9% 和 76.2%,结果表明,非局部操作与 scSE 模块结合可以增强特征时空上以及通道间的信息提高准确率,验证了 SC_NLResNet 网络的有效性。

关键词 双流卷积神经网络 scSE 模块 残差网络 非局部操作 行为识别

中图分类号 TP391.41 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.08.046

ACTION RECOGNITION ALGORITHM FOR NON-LOCAL TWO-STREAM RESNET NETWORK BASED ON SCSE FUSION

Li Zhanli Wang Jiaying Jin Hongmei Li Hongan
(*College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710600, Shaanxi, China*)

Abstract Aimed at the problem of low recognition rate of video frames containing redundant information in dual-stream network, scSE (Spatial and Channel Squeeze & Excitation Block) and non-local operation are introduced based on two-stream network to construct SC_NLResNet behavior recognition framework. In this framework, the framework divided the video into equal and non-overlapping temporal segments and sparsely sampled each segment, extracting RGB frames and optical flow graphs as the input of the scSE module. The features processed by scSE were inputted into the non-local two-stream ResNet network, and the segmentations were merged to obtain the final prediction results. The experimental accuracy on UCF101 and Hmdb51 dataset reaches 96.9 % and 76.2 %, respectively. The results show that the combination of non-local operation and scSE module can enhance the information of feature space-time and between the channels to improve the accuracy, which verifies the effectiveness of SC_NLResNet network.

Keywords Two-stream convolutional neural network ScSE module Residual neural network Non-local operation Action recognition

0 引 言

互联网的发展使视频数据迎来指数级的增长,各种摄像头、智能手机端等设备使得数据获取更加便捷。在现实生活中和实际的应用场景下,直接收集的视频还包含很多与行为无关的信息以及噪声,这些信息都对识别结果有一定影响。

卷积神经网络(Convolutional Neural Network,CNN)的研究,使得深度学习方法在行为识别领域迅猛发展,并取得一定成绩。卷积神经网络[1]的特征提取都是针对视频帧进行空间特征提取,而视频中的行为识别是基于连续的动作序列,其相邻帧之间具有行为关联,因此单一的卷积神经网络难以完成人体行为识别。孙月

收稿日期:2021 - 02 - 27。陕西省自然科学基础研究计划项目(2019JM-348,2019JLM-10)。**李占利**,教授,主研领域:计算机图形图像处理。**王佳莹**,硕士生。**靳红梅**,讲师。**李洪安**,副教授。

驰等[2]通过构建多层感知器层优化神经网络,增强卷积对前景目标特征提取的能力,未考虑到时序上的特征关联。Hochreiter 等[3]提出用长短时记忆(Long Short-Term Memory,LSTM)网络对时间维度进行建模,来获取时间特征。Gammulle 等[4]提出双流 LSTM 网络,并将其用于行为识别。但是长短时记忆网络不能充分提取视频长距离的关联信息。针对此问题王毅等[5]提出一种基于时空双流融合网络与注意力机制结合的方法来提高 LSTM 对视频信息提取不充分导致的识别率较低的问题。Ji 等[6]将卷积神经网络扩展到三维提出 3D 卷积神经网络,使用 3D 卷积可以提取时间以及空间的特征。但是相比于二维卷积神经网络,需要大量参数因而增大了网络训练的难度。Simonyan 等[7]提出双流卷积神经网络,将 RGB 视频帧与光流图像分别作为空间流与时间流的输入,通过空间流获得图像的空间信息,通过时间流获取视频的时域信息。但是双流网络不能充分利用时间维度信息,为了解决此问题,陈颖等[8]提出一种基于 3D 双流卷积和门控循环单元网络相结合的行为识别模型,利用视频中的时间维度的信息进行识别,但不能长时间保存信息。Wang 等[9]提出时序分割网络(Temporal Segment Network,TSN)改进了传统双流 CNN 的对长时序视频建模不足的问题,但是该方法忽略了重要特征的筛选。

基于以上分析,行为识别存在以下问题:(1) 循环神经网络和 LSTM 等改进方法捕获时序关联,以及 CNN 卷积捕获空间信息的方法都是局部信息操作,具有一定的局限性。(2) 双流网络忽略对输入图像的处理,没有考虑到图像中噪声等对行为识别的干扰。

针对以上问题,在 ResNet 网络基础上融合 scSE 模块,并采用非局部操作搭建双流网络,提出融合 scSE 的非局部双流 ResNet 网络,构建 SC_NLResNet 网络模型。非局部操作关注全局信息,捕获视频帧中远距离信息关联。scSE 模块关注对识别结果影响较大的重要特征特征,减少无用信息以及部分噪声的干扰,且解决了非局部操作忽略通道间信息的问题。二者结合能够增强关注时空信息,提高行为识别的准确率。

1 整体模型设计

1.1 模型框架

本文提出整体网络框架。针对传统双流网络无法关注图像重要特征的问题,采用 scSE[10]模块对重要特征赋予更多的关注度。并考虑到卷积和池化操作会对信息造成损失,不能获取更全面的信息,具有一定局限性,采用非局部操作获取全局信息减少信息丢失。同时 scSE 模块对非局部操作进行互补,两者结合可以关注到空间信息、时间信息以及通道之间的关系。搭建整体网络框架如图 1 所示。

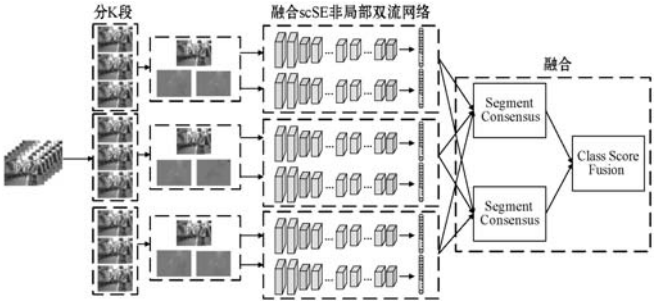


图 1 整体模型框架

由图 1 整体模型框架可知,先将视频在时序上初始划分为 K 等段,在每一段上进行稀疏采样,抽取部分 RGB 帧以及光流图以减少计算量。然后将视频稀疏采样得到的 RGB 帧以及光流图分别通过 scSE 模块进行特征筛选,关注对识别结果影响较大的特征。并将筛选出的重要信息分别作为非局部双流 ResNet 网络中时间流和空间流的输入。将相同通道不同段的信息进行融合,最终结合两流的信息得到预测结果。

1.2 非局部模块

图像中含有较多重复冗余信息,非局部均值(Non-Local Means,NLM)[11]可以利用图像中的冗余信息,在去噪的同时较大程度地保持图像的细节特征。

Wang 等[12]基于图片滤波领域的非局部均值滤波操作思想,提出了一个非局部操作算子,可以捕获时序信息和空间信息。非局部模块的结构如图 2 所示,其中:X 表示输入的图片或者特征,⊗表示矩阵乘法,⊕表示按照元素求和,θ 以及 ϕ 表示两个嵌入式高斯函数,g 是一个一元函数,Z 为输出图片或者特征。

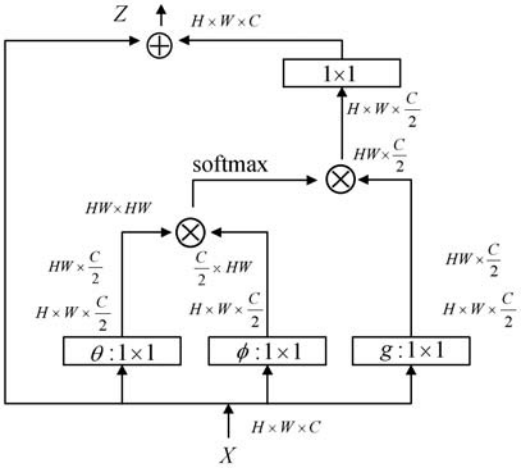


图 2 非局部模块结构图

首先输入维度为 $H \times W \times C$ 的图片,由于输入形式为图片,所以采用 1×1 的卷积核进行处理,将通道数

减半维度变为 $H\times W\times\frac{C}{2}$。通过 θ 和 ϕ 进行处理再经过 reshape 重排,将维度转换为 $HW\times\frac{C}{2}$和$\frac{C}{2}\times HW$ 的矩阵相乘得到 $HW\times HW$ 的矩阵,然后经过 softmax 进行归一化处理。即将当前特征图中的像素与其他位置像素计算相关性,并进行归一化操作。将经过 g 处理再进行 reshape 后的矩阵与 softmax 归一化处理后的矩阵进行矩阵乘法,得到维度为 $H\times W\times\frac{C}{2}$的矩阵,最后再经过一个 1×1 的卷积,将通道扩展为原来的通道维度,最终输出 Z 变为与原始输入 X 相同的通道维度。其输出为关联位置像素加权平均值,以此来关联远处像素的信息。

2　SC_NLResNet 网络模型

在时序的视频中,通常采用 RNN 和 CNN 卷积分别捕获长时序的关联以及空间上的信息,这些都是局部的信息操作。但是在图像或者视频中像素点与像素点间存在某种关联,并且相似像素并不局限于某个局部的区域。因此想要关联更全更远位置的信息时,这些局部操作存在局限性,故而采用非局部操作来处理视频帧中长距离的依赖关系。

非局部操作主要是寻找特征图中各像素点之间的关系,但忽略了特征图通道之间的信息。scSE 模块可以弥补非局部操作在通道间关注度的不足,同时对噪声的干扰有一定的鲁棒性。scSE 与非局部操作结合可以在时间、空间以及通道间均考虑到特征的关联性和重要性。

本文将 scSE 模块以及非局部操作进行结合,同时提出 SC_NLResNet 行为识别网络模型设计,如图 3 所示。

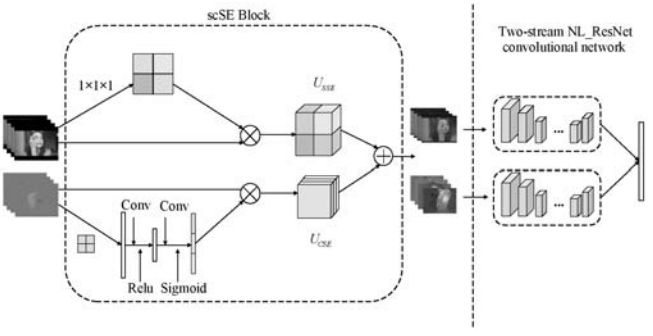


图 3　SC_NLResNet 网络模型设计

将视频稀疏采样得到的 RGB 帧以及光流图分别通过 scSE 模块进行特征筛选,得到处理后的特征。并将筛选出的特征分别输入基于 ResNet101 框架的非局部双流网络中时间流和空间流,对应图 3 中 NL_ResNet,然后融合两流的信息作为识别结果。

由图 3 中 scSE Block 可以看出,scSE 由 CSE(Spatial Squeeze and Channel Excitation Block)和 SSE(Channel Squeeze and Spatial Excitation Block)模块组成,直接接收稀疏采样得到的图片作为输入。其中 CSE 方法是将 $C\times H\times W$ 大小的 feature map 通过 global average pooling 方法变为 $C\times1\times1$ 大小。然后用卷积进行处理,再用 sigmoid 函数进行归一化,最后特征通道与权重相乘得到最终的特征图。SSE 是直接在 feature map 上进行 $1\times1\times1$ 的卷积,将 feature map 从 $C\times H\times W$ 变为 $1\times H\times W$ 大小,然后使用 sigmoid 函数,最后得到空间重新校准特征。

此过程为 scSE 处理阶段,scSE 模块同时关注通道以及空间信息,寻找特征通道之间关系以及特征空间上的关系。通过训练学习获取特征的重要程度,提升有价值特征的作用力,抑制无用或对结果影响小的特征。然后将处理后的特征作为非局部双流 ResNet 网络的输入,即图 3 中 NL_ResNet 网络中进行识别。

由于双流网络中空间流和时间流结构相同,图 4 只展示单流 SC_NLResNet 网络结构图。

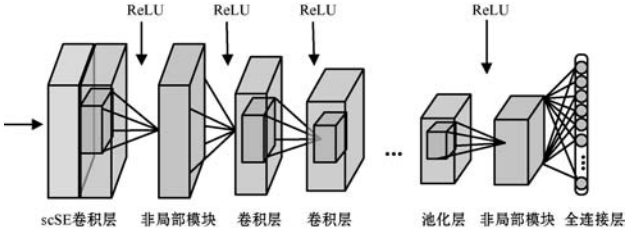


图 4　单流 SC_NLResNet 网络结构图

图 4 中,将 scSE 与第一个卷积层结合,称为 scSE 卷积层。非局部操作可以获取到全局的信息,因此与卷积操作互补进而两者可以有更好的效果,但非局部操作硬件消耗较大,为了平衡速度与准确率,采用两个非局部模块。一个非局部模块在第一个卷积层后,网络的浅层加入非局部操作可以更好地关注到特征,对图像具有一定的去噪功能,同时减少特征损失。另一个在全连接层前以获取全部信息来用于识别分类。

本文的网络框架由 scSE 模块、非局部模块以及主干网络 ResNet101 的双流网络组成,输入为 RGB 视频帧、光流图,具体网络处理流程如下。

(1) 分段。将视频划分为等长不重叠的时序视频段,对每一段进行稀疏采样,获得相应段的 RGB 帧以及光流图。

(2) 特征筛选。将采样得到的图作为 scSE 卷积层的输入。scSE 关注通道之间以及空间之间的关系,

对特征进行筛选,得到一个与输入大小相同的特征图作为卷积层的输入。

(3) 获取全局信息。然后将卷积处理后的特征作为非局部模块的输入来捕获特征图时空上的依赖关系,减少特征信息损失。

(4) 识别。将非局部操作处理后的特征,输入到后续 ResNet101 网络模型框架进行特征处理。当经过 RseNet 最后一个平均池化时,不直接将信息送入全连接层进行识别分类,而是将经过平均池化后的特征图作为第二个非局部模块的输入进行特征关联,然后再将非局部的特征信息作为全连接层的输入进行识别。

(5) 融合。将不同段相同流的结果融合,最终将时间流以及空间流的结果融合得到最终的预测结果。

3 实 验

3.1 数据集

视频人体行为识别领域常用数据集包括 KTH、Weizmann、UCF101,及 Hmdb51 等。为了验证本文提出的网络能更好地识别包含复杂背景的视频帧,同时为了比较数据集中视频信息的复杂程度,表 1 将不同数据集的复杂程度进行量化对比。

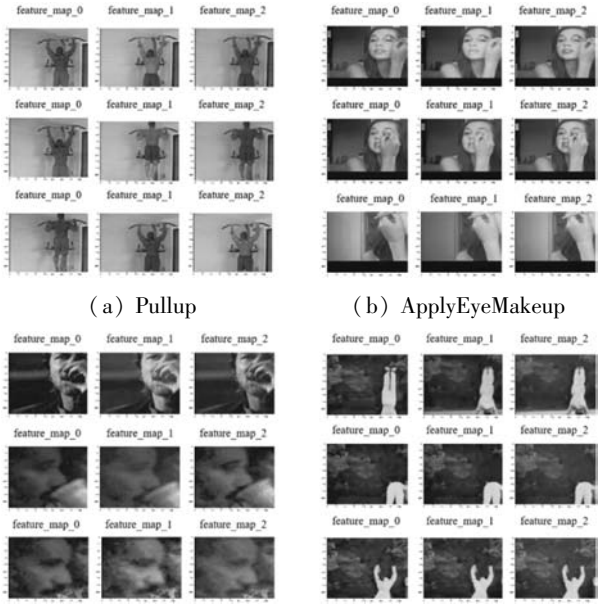
表 1 常用数据集

数据集	视频数	分类数	介绍
KTH	2 391	6	包含慢跑、拍手、拳击、步行等 6 类行为,且视频背景较为单一,摄像机静止拍摄
Weizmann	93	10	由 9 个人完成跑、跳、弯腰等 10 种动作,且相机拍摄角度相对固定,视频背景相对简单
UCF101	13 320	101	来源于 YouTube,包含人与人的交互、肢体运动等 101 类动作行为,具有较多的视频变化,更贴合现实场景
Hmdb51	6 849	51	部分来源于电影片段和 YouTube 等,包含人物交互,个人行为动作以及表情等,视频背景相对复杂,行为多样

可以看出,UCF101 以及 Hmdb51 数据集中的视频存在视点变化、光照变化、相机运动以及背景噪声等行为场景。本文以此数据集来验证算法在具有复杂背景的视频场景下的有效性。

3.2 特征可视化分析

scSE 模块关注空间上的特征信息将重要信息进一步强化,弱化对识别结果影响较小的特征信息以及噪声的干扰。本文将 scSE 模块与第一个卷积层结合称为 scSE 卷积层,放置在网络输入层的下一层,先对输入的特征进行筛选以及判断,然后将筛选后的特征根据关键程度赋予不同的权重值,输入到后面的网络层中进行识别。在数据集 UCF101 以及 Hmdb51 上进行可视化,RGB 作为输入时 scSE 模块处理后的特征图如图 5 所示。



(a) Pullup (b) ApplyEyeMakeup

(c) Drink (d) flic_flac

图 5 RGB 经过 scSE 模块处理后的特征图

在 UCF101 和 Hmdb51 数据集上进行实验,将训练过程中经过 scSE 模块处理后的特征提取如图 5 所示,其中图 5(a)、(c)、(d)分别为 Hmdb51 数据集中 Pullup、Drink、flic_flac 行为类别,图 5(b)为 UCF101 数据集中 ApplyEyeMakeup 类别。每个类别图中包含三行,表示列举的每类中的三组行为,每组即为图中一行,每行表示一次提取的 3 幅特征图,分别为 feature_map_0、feature_map_1 以及 feature_map_2。可以看出经过 scSE 处理的特征在网络的浅层仍然具有较好的信息表示。

图 5(a)Pullup 行为,可以看出经过 scSE 模块处理的特征依然可以保持较好的轮廓特征以及重要信息。并且图 5(b)ApplyEyeMakeup 动作可以直接从图中看出来,不会存在过大的信息损失现象。图 5(c) Drink 动作图中大部分信息为人脸信息,但识别重点信息是手和杯子,而处理后的特征图能够更好地抓住关键信息。图 5(d) flic_flac,可以看出关注的重点在人身上,并且对人的行为信息赋予更高的权重,而背景等大部分信息被弱化。同样在 Hmdb51 数据集上提取光流图如图 6 所示,将光流图作为网络的输入经过 scSE 模块进行处理,并将其进行特征可视化如图 7 所示。

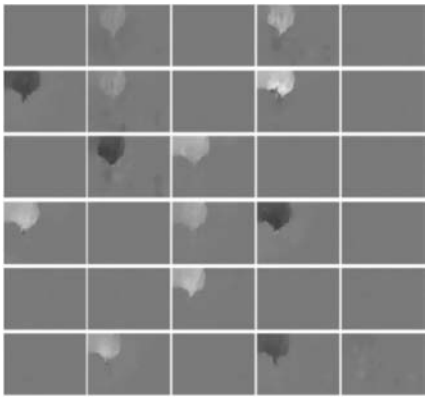


图 6　Hmdb51 数据集的光流图

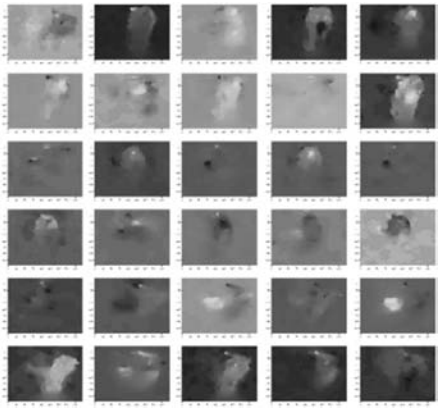


图 7　经 scSE 模块处理后光流图的特征图

图 6 为 TV-L1 方法提取 x 轴和 y 轴两个方向的光流,采集到的部分光流图。由于在行为视频中,某些行为移动只发生在水平方向,则水平方向采集的光流图更为清晰,垂直方向上位移变化不明显则生成的垂直方向光流图不清晰。反之,在垂直方向采集的光流更为清晰,水平方向上采集的光流不明显。可以看到,部分光流图在没有放大的情况下轮廓不明显,只有少量的光流图可以看到光流变化。而 Hmdb51 数据集上提取的光流图作为输入经过 scSE 处理并可视化显示如图 7 可知,经过 scSE 处理的光流图比图 6 没有处理的光流特征图更为清晰,且光流变化区域更加明显,因此经过处理后的特征更能表达运动行为的变化情况。

3.3　实验参数及其参数设置

深度学习实验环境:实验在 Windows 10 的 64 位系统下,Intel(R) Core(TM) i9-9820X CPU,GPU 2080 Ti ,32 GB RAM,运用深度学习框架 PyTorch,神经网络的学习测试环境 Python 3.6,NVIDIA CUDA 10.1。

初始学习率为 0.001,num_segments 表示视频在时序上初始分段数,即图 1 整体模型框架中的 K 段,经过实验验证在 UCF101 数据集上初始分段数 K 为 5,以及在 Hmdb51 数据集上初始分段数 K 为 7 识别效果最好。lr_steps 的值表示到达一定 epochs 的时候需要改变学习率,并设定将学习率修改成当前的 0.1 倍。输入 RGB 图像时,lr_steps 为 30 和 45,即 epochs 到达 30 时更改学习率为初始学习率的 0.1 倍,到达 45 时更改学习率为上次学习率的 0.1 倍。输入光流图时,lr_steps 等于 60,同样 epochs 等于 60 时更改学习率为原来的 0.1 倍。dropout 表示按照一定的概率将神经网络单元从网络中暂时丢弃,实验设置 dropout_rgb 为 0.8,dropout_flow 为 0.7。b 是 batch size,根据实验的硬件要求以及实验效果将其设定为 16。参数设置如表 2 所示。

表 2　实验参数设置

参数	UCF101		HMDB51	
	RGB	Optical Flow	RGB	Optical Flow
初始学习率	0.001	0.001	0.001	0.001
num_segments	5	5	7	7
Epochs	50	80	50	80
dropout	0.8	0.7	0.8	0.7
lr_steps	30,45	60	30,45	60
b	16	16	16	16

3.4　网络模型对比实验

本文在 UCF101 以及 Hmdb51 公共数据集上进行实验,且将数据集划分为三等分:数据集 1、数据集 2、数据集 3,每一份数据集又分为三部分:训练集,验证集、测试集,比例为 8∶1∶1,最终识别率为三份数据集识别结果的平均值,比较不同的网络框架对识别率的影响,实验效果如图 8 – 图 9 所示。

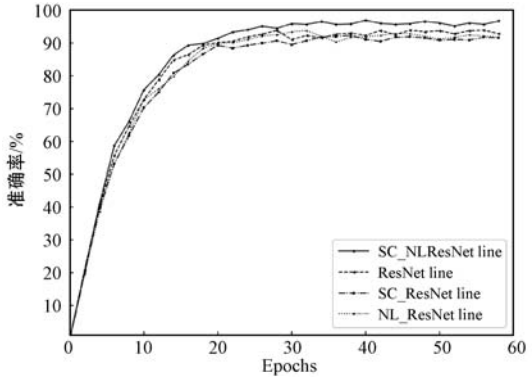


图 8　不同模型在 UCF101 数据集上对比实验

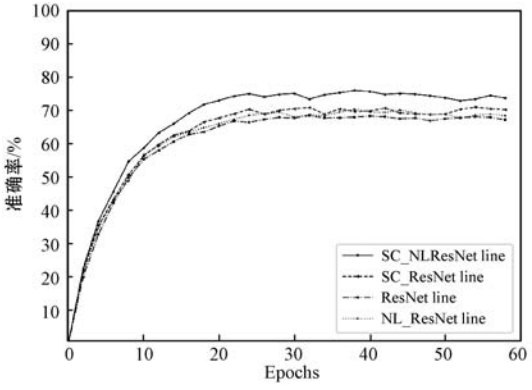


图 9　不同模型在 Hmdb51 数据集上对比实验

在 UCF101 数据集上分别采用四种模型：基于 ResNet 的双流网络框架、基于融合 scSE 的双流 ResNet 网络框架、非局部双流 ResNet 网络以及融合 scSE 非局部双流 ResNet 网络框架分别对应图 8 中 ResNet Line，SC_ResNet Line，NL_ResNet Line 以及 SC_NLResNet Line。可以看出，随着 Epochs 的增长，识别率不断提高，当 Epochs 达到 20 以上时，识别率基本趋于稳定且最终识别率均达到 85% 以上。

在 Hmdb51 数据集上进行实验，同样采用基于 ResNet 双流网络框架、基于融合 scSE 的双流 ResNet 网络框架、非局部双流 ResNet 网络以及融合 scSE 非局部双流 ResNet 网络框架，分别对应图 9 中 ResNet Line，SC_ResNet Line，NL_ResNet Line 以及 SC_NLResNet Line 进行对比。由于 Hmdb51 数据集具有较多背景复杂的视频，因此整体的识别率低于在 UCF101 数据集上的实验结果。从图 9 可以看出，随着 Epochs 的增长识别率不断提高。当 Epochs 达到 20 以上时，识别率基本趋于稳定，且识别率保持在 60% 到 80% 之间。同时为了探究不同模型的运行速度和准确率，表 3 分析了在相同实验环境下不同算法的准确率以及速度情况。

表 3 不同网络模型速度和准确率对比

算法	UCF101		HMDB51	
	准确率/%	平均耗时/ms	准确率/%	平均耗时/ms
Two-Stream	85.9%	172	57.6%	180
LTC	90.1%	185	63.4%	183
TSN(ResNet)	93.4%	530	68.3%	541
Ours(SC_ResNet)	95.8%	604	69.9%	599
Ours(SC_ResNet_NL_2)	96.3%	581	74.7%	597
Ours(SC_NLResNet)	96.9%	937	76.2%	924

在 UCF101 数据集以及 Hmdb51 数据集上进行实验如表 3 所示。为了更明确清晰地判断出模型运行的准确率以及运行速度变化情况，对提出的模型逐步增加网络层进行分析。表 3 中，SC_ResNet 网络为在双流 ResNet 网络基础上增加 scSE 卷积层，而 SC_ResNet_NL_2 网络为在 SC_ResNet 此基础上，增加全连接层前的第二个非局部层，SC_NLResNet 网络为在 SC_ResNet_NL_2 基础上，在 scSE 卷积层后增加第一个非局部层。

可以看出，相比于 Two-Steam 和 LTC 算法，TSN 网络在提高了识别准确率的同时耗费了一定的时间。与 TSN 方法相比，融合 scSE 模块的 SC_ResNet 网络在保证不耗费过多时间的情况下准确率在 UCF101 数据集以及 Hmdb51 数据集上分别提高了 2.4 百分点和 1.6 百分点。SC_ResNet_NL_2 在 UCF101 数据集和 Hmdb51 数据集上相比于 TSN 算法准确率提高 2.9 百分点和 6.4 百分点，且没有增加过多时间消耗。说明在全连接层前增加非局部操作捕获全局信息有助于最终对行为类别的判断。SC_NLResNet 算法，增加两个非局部操作，一个在第一层卷积后，另一个在全连接层前，在提升行为识别准确率的同时时间消耗较多。由于在网络的浅层，图像维度较大，没有经过过多的卷积池化进行信息压缩，因此关注全局的信息计算量将会增多。

实验显示，SC_ResNet 网络比传统双流网络以及基于 ResNet 的 TSN 网络的识别率都高，说明 scSE 模块将图片进行通道以及空间特征筛选可以减小无关特征对识别结果的影响，关注对识别影响大的重要特征上，提高网络在信息上的关注程度，从而提升网络的识别率。SC_NLResNet 网络相比于 SC_ResNet 网络和 SC_ResNet_NL_2 网络行为识别准确率最高，但是运行耗费的时间最长，说明增加非局部操作关注全局信息可以提高网络识别的准确率，但非局部操作在网络浅层时间消耗较大。因此要合理平衡识别率以及速度之间的关系。

3.5 实验效果对比

本文主要将提出 SC_NLResNet 方法与一些主流算法以及最近领先的算法进行比较。在 UCF101 数据集以及 Hmdb51 数据集上，表 4 列出双流网络、时序分割网络等算法与 SC_NLResNet 算法的比较。

表 4 不同数据集上行为识别算法准确率对比(%)

算法	UCF101	Hmdb51
Two Stream[7]	88.0	59.4
Two-stream fusion[13]	92.5	65.4
LTC[14]	91.7	64.8
TSN[9]	94.2	69.4
ARTNet[15]	94.3	70.9
Dynamic Image Networks + iDT[16]	96.0	74.9
本文方法	96.9	76.2

通过表 4 可以看出，相同算法在 UCF101 数据集上的识别率普遍比较高。进行特征融合的双流网络比传统双流网络在 UCF101 数据集以及 Hmdb51 数据集分别高出 4.5 百分点和 6 百分点，说明有效合理的特征融合有助于最后的分类。在 UCF101 以及 Hmdb51

数据集上,本文提出的 SC_NLResNet 算法比双流网络识别率分别高出 8.9 百分点和 16.8 百分点,比同样采用双流网络基础的 TSN 识别率分别高出 2.7 百分点和 6.8 百分点,识别率比 Dynamic Image Networks 与 iDT 结合的方法高 0.9 百分点和 1.3 百分点。说明在视频背景较为复杂的情况下,非局部模块与 scSE 模块互补可以关注到特征空间以及通道间的信息,对特征的筛选对提升识别率有较大的作用。

4 结　语

本文在视频背景复杂的场景下对多个样本行为进行人体行为识别。针对图像中信息关注度以及处理视频中长距离依赖的问题,提出 SC_NLResNet 网络。在数据集 UCF101 和数据集 Hmdb51 上进行实验,与传统的双流网络、TSN 等算法相比,SC_NLResNet 网络具有更好的识别率。双流 ResNet 网络使得网络在空间流以及时间流都获取到相应的视频信息,非局部操作获得全局的信息减少信息损失,使得网络能够获得更全面的特征信息。并且 scSE 弥补了非局部操作未关注通道信息的缺点,同时减少视频帧中噪声的影响,关注重要特征从而提升网络的识别率。但是网络的运行速度还达不到实时,今后的工作中将重点放在保证准确率的前提下对网络的实时性做更进一步的研究。

参 考 文 献

[1] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning,2009,2(1):1 – 127.

[2] 孙月驰,平伟,徐明磊. 基于优化卷积神经网络结构的人体行为识别[J]. 计算机应用与软件,2021,38(2):198 – 204,269.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation,1997,9(8):1735 – 1780.

[4] Gammulle H, Denman S, Sridharan S, et al. Two stream LSTM: A deep fusion framework for human action recognition[C]//IEEE Winter Conference on Applications of Computer Vision,2017:177 – 186.

[5] 王毅,马翠红,毛志强. 基于时空双流融合网络与 Attention 模型的行为识别[J]. 计算机应用与软件,2020,37(8):156 – 159,193.

[6] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(1):221 – 231.

[7] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Neural Information Processing Systems,2014,1(4):568 – 576.

[8] 陈颖,来兴雪,周志全,等. 基于 3D 双流卷积神经网络和 GRU 网络的人体行为识别[J]. 计算机应用与软件,2020,37(5):164 – 168,218.

[9] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: Towards good practices for deep recognition[C]//European Conference on Computer Vision,2016:20 – 36.

[10] Roy A G, Navab N, Wachinger C, et al. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks[C]//21st International Conference on Medical Image Computing and Computer-Assisted Intervention,2018:421 – 429.

[11] Buades A, Coll B, Morel J M. A non-local algorithm for image denoising[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2005:60 – 65.

[12] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:7794 – 7803.

[13] Feichtenhofer C, Pinz A, Zisserman A, et al. Convolutional two-stream network fusion for video action recognition[C]//Computer Vision and Pattern Recognition, 2016: 1933 – 1941.

[14] Varol G, Laptev I, Shmid C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2018,40(6):1510 – 1517.

[15] Wang L M, Li W, Gool L V, et al. Appearance-and-relation networks for video classification[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018:1430 – 1439.

[16] Bilen H, Fernando B, Gavves E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2018,40(12):2799 – 2813.

(上接第 253 页)

[14] Zhang Z X, Liu Q J, Wang Y H. Road extraction by deep residual U-Net[J]. IEEE Geoscience and Remote Sensing Letters,2018,15(5):749 – 753.

[15] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB]. arXiv:1706.05587,2017.

[16] Demir I, Koperski K, Lindenbaum D, et al. DeepGlobe 2018: A challenge to parse the earth through satellite images[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops,2018:172 – 181.

[17] Kingma D, Ba J. Adam: A method for stochastic optimization[EB]. arXiv:1412.6980,2014.