doi:10.11887/j.cn.202303016 http://journal.nudt.edu.cn

利用多时间尺度卷积的视频行为识别网络*

陈西江[1],梁全恩[1],韩贤权[2],安 庆[3]
(1. 武汉理工大学 安全科学与应急管理学院,湖北 武汉 430070;2. 长江科学院,湖北 武汉 430010;
3. 武昌理工学院 人工智能学院,湖北 武汉 430223)

摘 要:基于2D的行为识别网络通常融合多张视频帧的分类结果识别不同的行为,但其在卷积过程中缺少对时空特征提取。针对该问题,基于时间位移模块(temporal shift module,TSM)的思想设计了一组多时间尺度卷积,包含不同设计的卷积核以提取融合不同时间尺度的时空信息。通过控制多时间尺度卷积嵌入ResNet50网络的位置及其模块的参数设置,寻找最优的基于多时间尺度卷积的行为识别网络。使用PyTorch深度学习框架训练模型,在大型开源数据集Something-Something v2上进行了实验研究。结果表明,基于多时间尺度卷积的行为识别网络对行为识别准确率达到了59.47%,优于TSM等网络。

关键词:行为识别;卷积神经网络;分类;残差神经网络;PyTorch
中图分类号:TP391.4 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2023)03-136-10

Video behavior recognition network using multi time-scale convolution

CHEN Xijiang[1], *LIANG Quanen*[1], *HAN Xianquan*[2], *AN Qing*[3]
(1. School of Safety Science and Emergency Management, Wuhan University of Technology, Wuhan 430070, China;
2. Changjiang River Scientific Research Institute, Wuhan 430010, China;
3. School of Artificial Intelligence, Wuchang University of Technology, Wuhan 430223, China)

Abstract: The behavior recognition network based on 2D convolutional usually integrates classification results of multiple video frames to recognize different behaviors, but it can't extract space-time feature using the 2D convolution kernels. To solve this problem, MTSC (multi time-scale convolution) was proposed based on TSM(temporal shift module), which contained convolution kernels of different scales to fuse the space-time feature from different time scales. By controlling the position that inserting MTSC into ResNet50 network and the parameter setting of MTSC, the optimal behavior recognition network based on MTSC was discussed. Using the PyTorch training model, an experimental study was conducted on a large open source dataset, Something-Something v2. The results show that the behavior recognition network based on MTSC achieves 59.47% Top-1 accuracy, and outperform TSM and other behavior recognition networks.

Keywords: behavior recognition; convolution neural network; classification; residual neural network; PyTorch

得益于计算机设备的进步与算力的提升,深度学习技术得到了快速发展。许多学者提出了基于卷积神经网络的图像识别算法,如:AlexNet[1]、VGG[2]、ResNet[3]等。由于神经网络在图像识别领域的优势,许多学者尝试运用卷积神经网络进行行为的识别与分类。基于不同的骨架网络,行为识别网络一般分为2D行为识别网络与3D行为识别网络。

2D的行为识别网络使用2D卷积神经网络作为骨架网络进行行为识别。Simonyan等[4]设计了包含两个独立卷积神经网络的双流网络,其以密集的连续帧作为网络输入提取时序信息。但是密集的连续帧无法对动作进行大时间尺度的建模。为改进这一缺点,Wang等设计了时间分割网络(temporal segment network,TSN)[5]。TSN将视频分段,将每段视频输入到双流网络中再对每段的结果进行融合从而使网络具有长时时空建模的能力。Zhou等提出时间关系网络(temporal relation network,TRN)[6]。TRN主要关注不同时间尺度上的不同帧的相关性,其将图像特征依照不同的时间尺度进行时间关系推理得到不同时间尺度下的行为分类结果,最后融合多尺度的分类

* **收稿日期**:2021-06-01
基金项目:国家自然科学基金资助项目(42171428);重庆市技术创新与应用发展专项面上资助项目(cstc2019jscx-msxmX0051);长江科学院开放研究基金资助项目(CKWV2019758/KY)
作者简介:陈西江(1985—),男,安徽淮南人,副教授,博士,硕士生导师,E-mail:cxj_0421@163.com

结果得到最终的分类结果。Zolfaghari 等提出了一种高效的行为识别网络[7],其创新在于在网络底部使用 3D 卷积神经网络来获得最后的分类结果。基于动作主体语义变化相较于动作变化本身更慢,Feichtenhofer 等设计了 SlowFast 网络[8],SlowFast 网络包含了两个不同设计的卷积神经网络,分别侧重于提取不同变化速率的特征。Yang 等[9]设计了一个金字塔结构的时间金字塔网络(temporal pyramid network,TPN)用以对动作的不同速率进行采样,其利用不同层次网络的输出特征,应用不同的空间采样率与时间采样率进行采样,最后将采样后的特征融合获得行为的分类结果。刘董经典等出了 2D 时空卷积密集连接神经网络[10]。他们选取视频中用于表征行为的帧,将这些帧依照不同的时空次序组成蓝绿红(blue green red,BGR)格式的数据,将组成的图片数据输入 2D 时空卷积密集连接神经网络以对行为进行识别分类。

3D 的行为识别网络利用 3D 卷积核构建的卷积神经网络,卷积核本身扩张了时间维度,从而在卷积过程直接提取输入图像间的时序信息。3D 卷积神经网络 C3D[11] 由 Tran 等首次提出用于行为识别。但 3D 卷积核扩展维度会使网络的参数量成倍增加。因此,Qiu 等提出了 Pseudo-3D 网络[12],P3D 网络将 3D 卷积核进行了分解以降低参数量。Tran 等提出了 R(2+1)D 网络[13],其思路与 P3D 网络的思路相似,但在分解卷积核时保持了参数量一致。张小俊等[14] 借鉴 P3D 网络,但相比直接替换卷积核,他们设计了一种双流的网络结构。Carreira 等设计了一个双流 3D 卷积神经网络 I3D[15],他们探讨了如何应用图像分类和识别模型的预训练参数于 3D 卷积神经网络中。Xie 等提出了 S3D 网络[16],S3D 在 I3D 网络的基础上对 I3D 网络内的 Inception block 中的 3D 卷积核进行分解。Qiu 等[17] 基于分组卷积设计了一个提取时空特征的卷积模块组分解模块(grouped decomposed module, GDM)并构建了行为识别网络组分解网络(grouped decomposed network, GDN)。GDM 将输入特征沿通道分为三部分,分别使用不同的卷积核计算,最后将计算结果沿通道拼接从而融合不同时空信息。郭明祥等提出三维残差稠密的行为识别网络[18]。他们将 DenseNet 中的卷积核替换为 3D 卷积核,利用网络本身的密集连接融合不同层级的时空特征,使用自适应的局部特征与全局聚合来学习行为的局部密集特征与全局特征。

基于 2D 卷积神经网络的行为识别网络在卷积过程缺少对时空特征的提取,因而限制了其性能。Lin 等提出了时间位移模块(temporal shift module,TSM)[19] 尝试解决 2D 的行为识别网络存在的问题。本文受 TSM 的启发提出了多时间尺度卷积。相比 TSM,本文提出的多时间尺度卷积能够更好地融合前后多帧的信息到当前帧中,使网络获得更好的时空建模能力。本文讨论了多时间尺度卷积的具体设计与其在骨架网络 ResNet50 中插入的位置与数量,构建了行为识别网络,并在大型开源数据集 Something-Something v2 上进行实验对比。

1　多时间尺度卷积设计与网络构建

1.1　卷积神经网络与残差神经网络

卷积神经网络由多个卷积层、池化层与全连接层组成。卷积层一般由卷积核和激活函数或其他组件组合而成。这些基础的组件以串联或并联的方式连接,输入的图像特征依照顺序送入每一个组件最后得到该卷积层的输出。卷积层计算式可以表达为:

$$O = f(W_{(\theta)}(x)) \tag{1}$$

式中,x 为卷积层输入,O 为卷积层输出,W 代表卷积核,f 为激活函数,θ 为卷积核参数。输入图像经过多个卷积层的计算被提取为高维特征,之后将高维特征展开以一维向量的形式输入到全连接层中得到分类结果。得益于卷积核强大的特征提取能力,卷积神经网络在多个数据集上的性能表现都超过了滑动窗口、手工特征、多层感知机等传统方法。同时卷积核共享参数的特性使得卷积神经网络计算更高效且易于训练。

残差神经网络 ResNet 是由 He 等提出的一系列卷积神经网络,其在多个开源数据集上取得了较高的分类准确率。ResNet 依照网络层数不同可以划分为 ResNet18、ResNet34、ResNet50 等网络。以 ResNet50 为例,如表 1 所示,其网络由 49 个卷积核和 1 个全连接层组成,依照不同输出特征大小,这些卷积核被分入不同的网络层。在网络层中,这些卷积核又被组织成瓶颈结构的形式。

瓶颈结构如图 1 所示,每个瓶颈结构包含参数为 1×1,3×3 和 1×1 的三个卷积核。两个大小为 1×1 的卷积核置于串联结构的顶部与底部,大小为 3×3 的卷积核置于结构的中部。瓶颈结构中 1×1 卷积核将输入特征的通道进行压缩与还原,3×3 卷积核在计算过程中保持通道数不变。瓶颈结构通过降低中间特征的通道数,显著

地减少网络的参数量并加快网络的训练速度。

表1 ResNet 50 结构

Tab. 1 Architecture of ResNet 50

结构	输出大小	卷积核参数
conv1	112×112	7×7
网络层1	56×56	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times3$
网络层2	28×28	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times4$
网络层3	14×14	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times6$
网络层4	7×7	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times3$
fc	1×1	

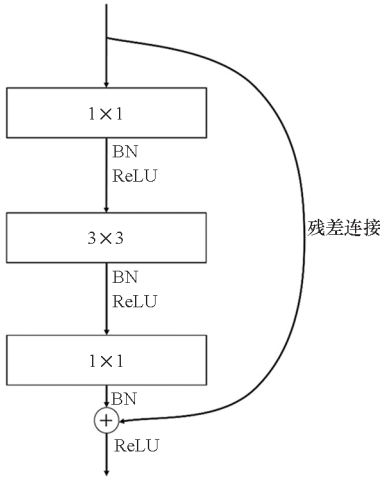

图1 瓶颈结构

Fig. 1 Bottleneck block

ResNet 在卷积核后使用激活函数 ReLU 和归一层 BN。BN 层通过归一化网络输出,加快网络训练的收敛速度,缓解梯度爆炸或梯度弥散的出现,同时抑制网络过拟合现象,因此 BN 层被广泛应用于各种神经网络中。BN 层的计算公式为:

$$\tilde{x}=\frac{x-\mu}{\sqrt{\sigma^2}}\cdot\gamma+\beta \tag{2}$$

式中:x 为输入数据;μ 为输入数据的均值;σ^2 为输入数据的方差;γ 与 β 为可训练参数参与到神经网络的训练过程,用于还原数据的分布。

激活函数 ReLU 将输入数据中小于0的数值置为0,大于0的数值保持不变,增加卷积神经网络的非线性因素。激活函数 ReLU 可以表述为:

$$\mathrm{ReLU}(x)=\max(0,x) \tag{3}$$

如图1所示,ResNet 的瓶颈结构引入了残差连接。通常,卷积神经网络的层数增加可以对输入特征进行更细致地拟合,但是随着网络层增加,网络变得难以训练,且其性能也不一定超越浅层网络。残差连接的提出有效地解决了深层网络的以上问题。设 x_{in} 为输入特征,x_{out} 为输出特征,$\varphi(x_{\mathrm{in}},\omega)$ 为卷积层代表的输入到输出的映射,其中 ω 为卷积运算。一个包含残差连接的瓶颈结构的计算过程可以表示为:

$$x_{\mathrm{out}}=x_{\mathrm{in}}+\varphi(x_{\mathrm{in}},\omega) \tag{4}$$

当映射 $\varphi(x_{\mathrm{in}},\omega)$ 的值逼近于0时有 $x_{\mathrm{in}}\approx x_{\mathrm{out}}$,此时认为该层网络没有学习到新的特征,即该层网络是输入到输出的一个近似的恒等映射。通过残差连接,使网络在层数增加时更易训练,并且维持网络性能不会退化。

表2是不同层数 ResNet 网络在 ImageNet[20] 数据集上进行图像分类的准确率。由表2可以看到,随着网络层数增加网络的识别准确率呈现上升趋势,说明残差连接有效地解决了前文提到深层网络存在的问题。

ResNet 系列网络结构简单,适合根据需求对其进行不同修改。通过对比表2中不同层数 ResNet 的准确率与参数量,选取在两者之间取得较好平衡的 ResNet50 作为本文的骨架网络。

表2 ResNet 在 ImageNet 数据集上的准确率

Tab. 2 Accuracy rate of ResNet on ImageNet dataset

模型	Top-1/%	Top-5/%	参数量
ResNet18	69.758	89.078	11.7×10^6
ResNet34	73.314	91.420	21.8×10^6
ResNet50	76.130	92.862	25.6×10^6
ResNet101	77.374	93.546	44.5×10^6
ResNet152	78.312	94.046	60.2×10^6

1.2 TSM

TSM 由 Lin 等提出。TSM 通过移动输入特征的部分通道将相邻两帧的部分特征引入当前帧中达到信息融合的目的,使骨架网络获得时空建模能力。TSM 的结构如图2所示,其中不同颜色的行对应不同时间点 T 的图像特征,两个箭头分别

为前向移动与后向移动。前向移动将部分特征沿着时间维度的顺序向前移动一个时间单位,通过前向移动每一帧都将融合后一帧的部分信息。后向移动沿着时间维度将部分特征向后移动一个时间单位,从而使每一帧获得前一帧的部分信息。T_0 与 T_4 的部分通道由于移动会出现数据缺失,TSM 中使用零值进行填充,对于超出时间范围的特征则舍去。TSM 使用 α 控制移动的通道数,通过参数控制,在不影响骨架网络空间建模能力的基础上最大限度地融合前后帧的信息。此外,TSM 可以不经修改原网络结构快速地插入到任意 ResNet 系列的网络中实现即插即用。

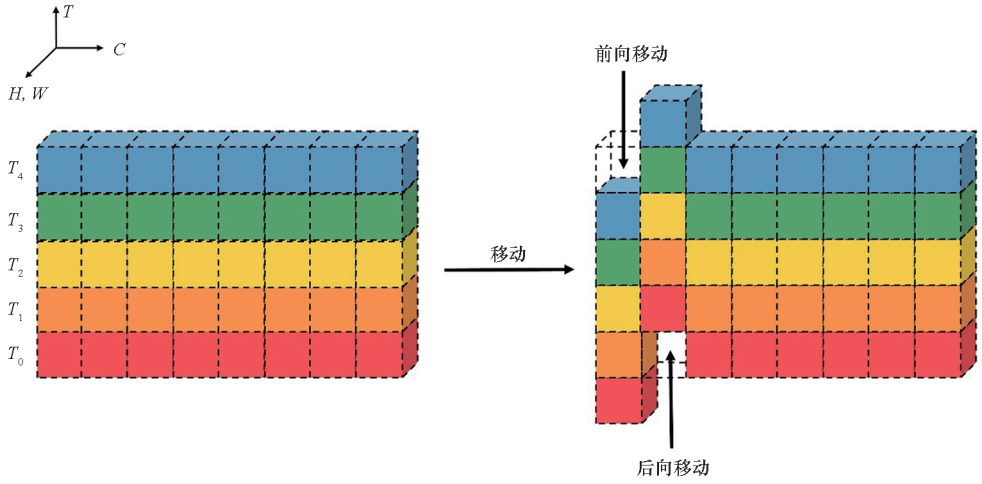

图 2　TSM 结构

Fig. 2　Architecture of TSM

TSM 的通道移动操作主要涉及数据在内存之中的移动,本文结合文献[19]说明 TSM 可以视作一个特殊的卷积模块。设输入特征 F 含有 n 帧图像,设 $t_1,t_2,\cdots,t_n$ 为 F 中不同帧的图像特征对应的时间点。将每一帧特征的通道分为三部分,F_{forward} 为 TSM 中需要前向移动的特征,F_{backward} 为需要后向移动的特征,F_{remain} 为不需要移动的特征。又设三个固定参数的 3D 卷积核为 c_1、c_2、c_3,将其时间维度的参数设为[0,0,1],[1,0,0],[0,1,0]。将 c_1、c_2 与 c_3 分别与 F_{forward}、F_{backward}、F_{remain} 进行卷积计算。以 c_1 与 F_{forward} 进行卷积计算为例,F'_{forward} 为输出特征,计算过程为:

$$\begin{cases} F'^{t_1}_{\text{forward}} = (0\times F^{t_0}_{\text{forward}} + 0\times F^{t_1}_{\text{forward}} + 1\times F^{t_2}_{\text{forward}}) = F^{t_2}_{\text{forward}} \\ F'^{t_2}_{\text{forward}} = (0\times F^{t_1}_{\text{forward}} + 0\times F^{t_2}_{\text{forward}} + 1\times F^{t_3}_{\text{forward}}) = F^{t_3}_{\text{forward}} \\ F'^{t_3}_{\text{forward}} = (0\times F^{t_2}_{\text{forward}} + 0\times F^{t_3}_{\text{forward}} + 1\times F^{t_4}_{\text{forward}}) = F^{t_4}_{\text{forward}} \\ \quad\vdots \\ F'^{t_n}_{\text{forward}} = (0\times F^{t_{n-1}}_{\text{forward}} + 0\times F^{t_n}_{\text{forward}} + 1\times F^{t_{n+1}}_{\text{forward}}) = F^{t_{n+1}}_{\text{forward}} = 0 \end{cases} \tag{5}$$

式(5)中的时间范围为 $t_1 \sim t_n$。该式中上标为 t_0 与 t_{n+1} 的 F 的值设为0,其为卷积过程中为维持特征大小不变所设置的参数。由式(5),与 c_1 计算后,F_{forward} 中当前时间点的特征变为了后一时间点的特征。同理可推 c_2 与 F_{backward},c_3 与 F_{remain} 相应的计算过程。经推导可知,通过固定卷积核时间维度上不同位置参数,可以使卷积舍去或保留不同时间点的特征,从而等价于不同的移动操作。综上所述,TSM 的移动过程可以表达为:

$$F' = C_{\text{Cat}}[c_1(F_{\text{forward}}), c_2(F_{\text{backward}}), c_3(F_{\text{remain}})] \tag{6}$$

式中,C_{Cat} 为拼接操作。由上述讨论,TSM 的通道移动操作可以视为使用不同的固定参数的 3D 卷积核与特征不同部分的通道进行卷积。与普通卷积核不同的是这些卷积核在训练过程中不学习参数。

1.3　多时间尺度卷积

TSM 证明了在 2D 骨架网络的基础上,使用部分输入特征进行信息融合可以使模型具有捕获时空信息的能力。受 TSM 的启发,本文设计了多时间尺度卷积(multi time-scale convolution,MTSC)提取融合帧间时空特征。多时间尺度卷积由两个时间 1D 卷积核组成:其一为大小 3×1×1 的时间 1D 卷积核用于提取当前帧及邻近前后两帧的特征,时间跨度为 3 帧;其二为大小 3×1×1 的空洞时间 1D 卷积核,用于提取当前帧及前后隔帧的特征,时间跨度为 5 帧。MTSC 的运算过程如图 3 所示,首先将原特征沿着通道顺序分割为截取特征 F_{conv} 与保留特征 F_{unconv},之后将截取特征分别输入到两个不同的时间 1D 卷积核

中进行计算以提取不同尺度的时空的信息,最后将卷积输出特征相加融合再依照通道顺序与保留特征拼接。如1.2小节讨论,TSM可以视作对特征的不同部分进行固定参数的卷积计算,但其存在两个缺点:参数不能学习;部分通道出现信息缺失。多时间尺度卷积使用可训练的时间1D卷积解决了以上两个问题,其表达式为:

$$F_{\text{out}} = C_{\text{Cat}}[k_1(F_{\text{conv}}) + k_2(F_{\text{conv}}), F_{\text{unconv}}] \quad (7)$$

式中,k_1,k_2 代表两个时间尺度的时间1D卷积。通过融合不同时间尺度的特征,输入中的每一帧特征获得前后不同时间尺度上的信息从而使网络具有更好的时空建模能力。

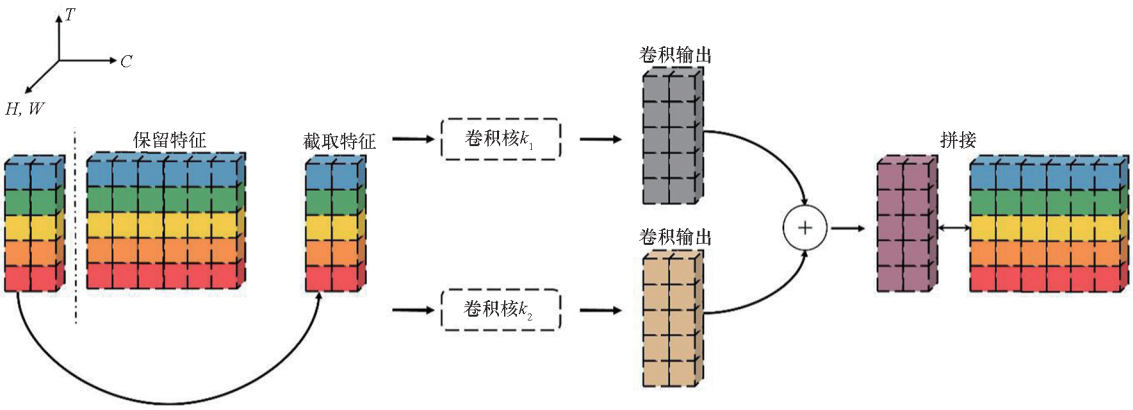

图3 多时间尺度卷积

Fig. 3 Multi time-scale convolution

1.4 行为识别网络设计

行为识别网络由若干个多时间尺度卷积嵌入ResNet50构成,因此本文提出的行为识别网络同时考虑了多时间尺度卷积结构与多时间尺度卷积的嵌入位置。

提出的多时间尺度卷积的结构为图3中截取特征的通道数量。截取特征通道数量不仅影响多时间尺度卷积提取融合时空特征的能力,并且随着通道数的增加,多时间尺度卷积的参数量也会上升。本文参考TSM,使用参数 α 来控制截取特征的通道数。α 代表输入特征总通道数 C_{in} 与截取特征通道数 C_{conv} 的比值。

$$\alpha = \frac{C_{\text{in}}}{C_{\text{conv}}} \quad (8)$$

多时间尺度卷积的嵌入位置指多时间尺度在骨架网络中具体嵌入的层数与数量。本文选取的骨架网络ResNet50含有多个瓶颈结构,多时间尺度卷积可以方便地嵌入到瓶颈结构之前。插入多时间尺度卷积的数量影响着网络的时空特征提取能力,并且嵌入多时间尺度卷积的数量也在影响模型的参数量,因此需要研究如何取得性能与参数之间的平衡。如表1所示,ResNet50包含网络层1至网络层4四个网络层,将多时间尺度卷积插入不同网络层的瓶颈结构前并进行对比,研究多时间尺度卷积在骨架网络中的最佳插入位置与数量。

本文提出的基于多时间尺度卷积的行为识别网络总体结构如图4所示。首先对视频进行稀疏采样,每个视频抽取8帧图像堆叠组成网络的输入。然后网络使用多个卷积层对输入图像进行特征提取。最后将卷积层输出的特征平铺为一维向量输入到fc层中,将fc层的输出相加并按帧数取均值得到识别结果。

2 实验与结果

2.1 网络性能评价指标与数据集

使用行为识别领域中常用的Top-1准确率与Top-5准确率作为性能评价指标。Top-1准确率是指网络的输出中概率最高的类别和视频实际类别一致的比例,Top-5准确率是指网络输出中概率最高的前五个类别中包含视频实际类别的比例。Top-1准确率与Top-5准确率的伪代码见算法1。

Something-Something v2数据集是一个大型的开源行为识别数据集。Something-Something v2数据集涵盖了174个行为类别,包括日常中常见的行为如:移动某物靠近某物、上移某物、打开某物等。Something-Something v2数据集中的动作类别注重时空上的关系,对于模型理解动作主客体之间的交互要求较高。Something-Something v2数据集共包含220 847个视频,其中训练集168 913个视频,测试集27 157个视频,验证集24 777个视频。为了在实验阶段快速验证网络性能,本文对

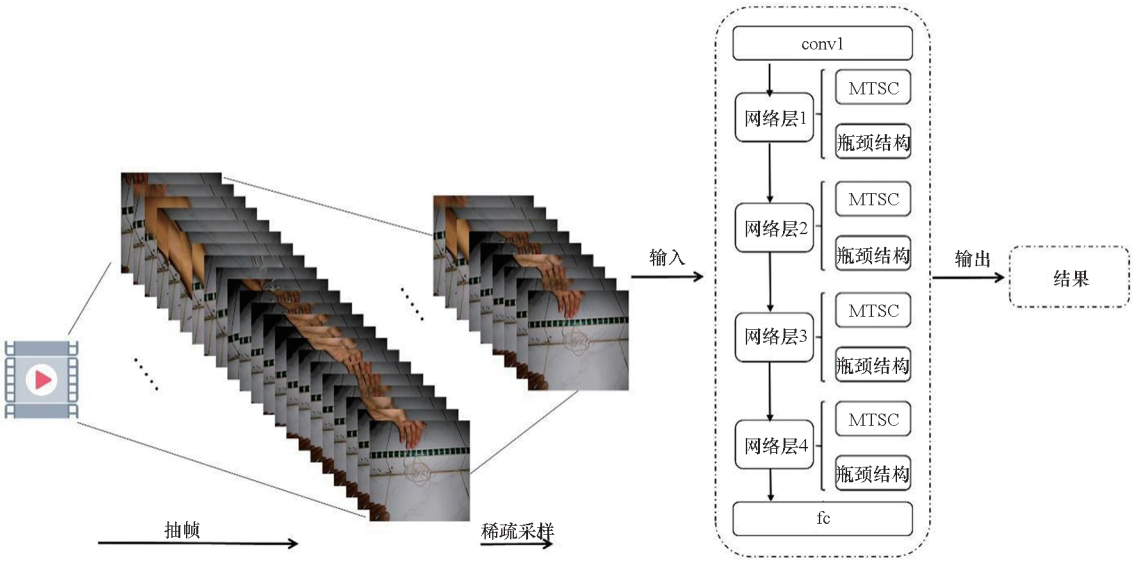

图 4　行为识别网络总体结构

Fig. 4　Overall of behavior recognition network

算法 1　Top－1 准确率与 Top－5 准确率

Alg. 1　Top－1 accuracy and Top－5 accuracy

输入: 卷积输出 O, 标签 L, 视频总数 N

输出: Top－1 准确率 Top1, Top－5 准确率 Top5

```
top1_correct = 0, top5_correct = 0
top5_possibility,top5_class = GetMax5Index(O)
top1_class = GetMax1Index(top5_possibility)
for each video, label ∈ N,L:
    if top1_class 等于 label:
        top1_correct + =1
    if label 包含于 top5_class 中:
        top5_correct + =1
end for each
Top1 = top1_correct/N
Top5 = top5_correct/N
return Top1,Top5
```

训练集中所有类行为的视频进行等比例选取,选取比例为 1/5,构成了包含 33 689 个视频的训练集子数据集(后文简称为训练子集),数据集划分情况如图 5 所示。

2.2　训练测试设置

实验环境为 Ubuntu16.04,一块 NVIDIA RTX 2080ti GPU,Pytorch 版本 1.4,Cuda 版本 10.0。由前文所述,选择 ResNet50 作为骨架网络并使用 ImageNet 预训练参数。本文选择 SGD 作为优化器,初始的学习率设置为 0.01,在第 20 和第 40 轮次时学习率下降为当前学习率的 1/10,优化器动量 momentum 为 0.8,模型训练的轮次为

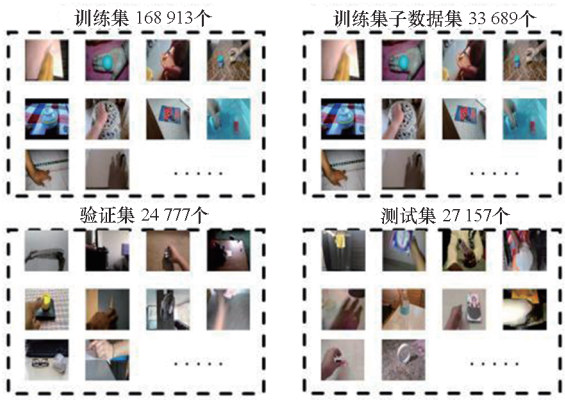

图 5　数据集划分

Fig. 5　Splits of dataset

50epoch。全连接层的 dropout 设置为 0.5。使用梯度累加将批大小模拟为 64。在训练时,在视频中抽取 8 帧的视频切片,将视频切片中的每一帧图像随机剪裁出 224×224 大小的图像,之后重新组成一个视频切片输入网络。在测试阶段与验证阶段,选取 8 帧视频切片,每一帧图像都在中心剪裁 224×224 大小的图像,之后重新堆叠输入网络进行测试。在实验部分,使用训练子集训练网络研究多时间尺度卷积的设计与卷积插入的层数选择。在进行与其他网络性能对比时,将使用完整的训练集训练网络。由于 Something-Something v2 数据集的测试集并未提供标签信息,因此将在提供标签信息的验证集上测试网络的性能。

2.3　实验

2.3.1　最优 α 值确定

参考 TSM 的研究,选取了三个 α 值分别为 2、4、8。在该实验中,多时间尺度卷积与 TSM 插入的

位置为网络层1至网络层4的瓶颈结构前。表3为在训练子集上对不同 α 值的网络进行训练并在验证集上测试的结果。α 值越小代表图3中截取特征的通道数越多。

表3 不同 α 值对应不同的网络精度

Tab. 3 Network accuracy by different α

模型	α	Top－1/%	Top－5/%
MTSC	2	42.94	71.70
MTSC	4	44.61	73.67
MTSC	8	43.87	72.30
TSM	4	43.87	72.69

表中 Top－1 与 Top－5 分别为 Top－1 准确率与 Top－5 准确率，在无其他说明的情况下后文中的其余表格与此相同。针对 Top－1 准确率，由表3可明显看出，$\alpha=4$ 时的网络精度相对 $\alpha=2$ 与 $\alpha=8$ 的网络分别提升1.67%和0.74%，同时比 TSM $\alpha=4$ 时提升0.74%。针对 Top－5 准确率，$\alpha=4$ 时的网络精度相对 $\alpha=2$ 与 $\alpha=8$ 的网络分别提升1.97%和1.37%，同样比 TSM $\alpha=4$ 时的网络精度提升了0.98%。因此，当 $\alpha=2$ 与 $\alpha=8$ 时，网络的性能都有不同程度的下降。根据文献[19]可以确定，造成该现象的原因如下：当 $\alpha=2$ 时，输入特征每一帧的特征都只保留了一半的原特征，这造成了较为严重的信息丢失，因此损害了网络的空间建模能力，进而导致网络性能下降。当 $\alpha=8$ 时，虽然保留了输入特征的绝大部分特征，但是时序信息融合较少，因此网络性能仍有上升空间。从表3可以看到，$\alpha=8$ 时多时间尺度卷积的性能与 TSM $\alpha=4$ 时的性能相似，证明了多尺度卷积比 TSM 具有更好的时空信息提取融合能力，能使用较少的通道数达到 TSM 中移动较多通道的效果。通过该实验可以确定 $\alpha=4$ 时网络取得了最优性能，因此，多时间尺度卷积的最优 α 值为4。

图6展示了部分行为类别在不同 α 值下的识别情况。由图6可以看到在 α 为2时，网络对某些类别的识别正确率下降严重，如“扭某物”。“扭某物”类别对空间信息较为敏感，因此可以印证前文的推测即 α 取值过大导致网络空间建模能力下降。

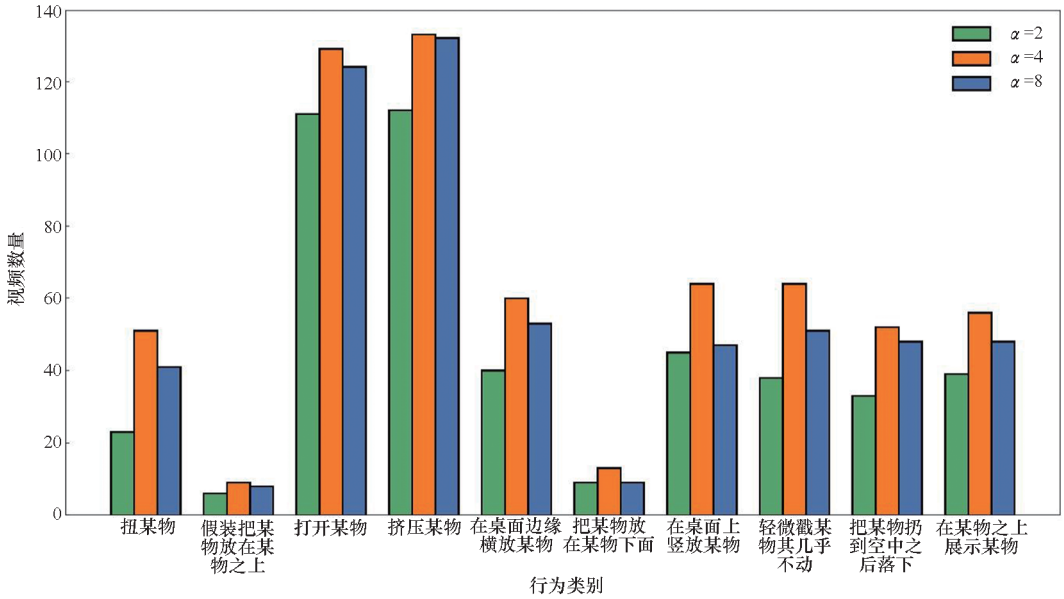

图6 不同 α 对应的部分类别识别结果

Fig. 6 Classification result of some behavior categories by different α setting

2.3.2 多时间尺度卷积插入位置确定

多时间尺度卷积可以简单地插入骨架网络中，将多时间尺度卷积分别在不同层内的瓶颈结构之前插入。选择了3种插入位置组合：[1,2,3,4]、[2,3,4]、[3,4]。[1,2,3,4]代表在第1、2、3、4层的每一个瓶颈结构前插入多时间尺度卷积，其余以此类推。在该实验中，多时间尺度卷积与 TSM 的 $\alpha=4$。实验结果见表4。

表4 不同插入位置的网络精度

Tab. 4 Network accuracy by different insertion position

模型	位置	Top－1/%	Top－5/%
MTSC	[1,2,3,4]	44.61	73.67
MTSC	[2,3,4]	44.29	73.19
MTSC	[3,4]	44.08	72.35
TSM	[1,2,3,4]	43.87	72.69

可以看到,随着插入层数的减少网络性能呈现下降趋势,这说明卷积插入数量的提升可以显著地增强网络的时空建模能力。但随着卷积核数量增加,网络的参数和计算消耗也会增大,因此对于部分计算量敏感的应用场景可以选择插入较少的层次如[2,3,4]。图 7 展示了不同插入位置对应的部分类别识别结果,说明随着插入层数的增加即插入多时间尺度卷积的数量增加,有利于模型的识别性能提高。

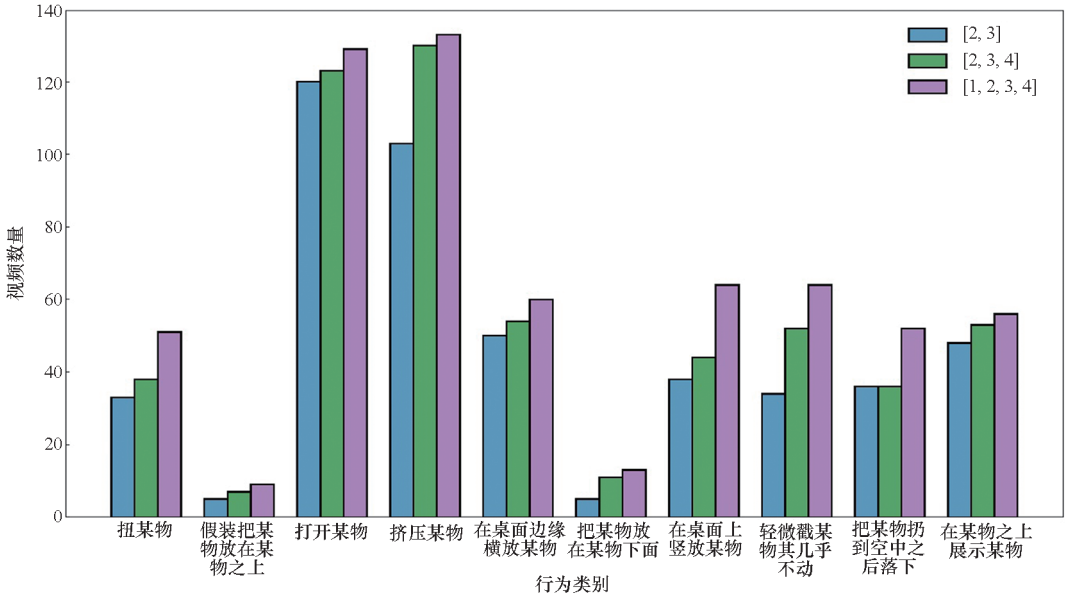

图 7　不同插入位置对应的部分类别识别结果

Fig. 7　Classification result of some behavior categories by different insertion position setting

2.4　模型性能对比

通过 2.3.1 节与 2.3.2 节的讨论,行为识别网络在多时间尺度卷积的 $\alpha=4$、插入层次为[1,2,3,4]时取得最好的性能。本节利用 Something-Something v2 数据集验证本文提出的网络与 TSN、TRN、TRN-2Stream 等网络的性能,使用 Top－1 和 Top－5 准确率对不同方法性能进行比较,结果见表 5。

表 5　与其他模型的对比

Tab. 5　Compare with other models

模型	帧	大小	Top－1/%	Top－5/%
TSN	16	224	30.00	
TRN	8	224	48.80	77.60
TRN-2Stream	8	224	55.50	83.10
TSM	8	224	58.70	85.47
TSN＋TPN	8	224	55.20	
GDN	8	224	57.60	84.60
	16	224	59.20	85.10
MTSC	8	224	59.47	85.54

TSN 为早期方法,其使用 16 帧图像作为输入仅取得了 30% 的 Top－1 准确率,落后于其他行为识别模型。由表 5 可以看出,针对 Top－1 准确率,基于多时间尺度卷积的行为识别网络超过了 TRN 以及使用光流输入的 TRN-2Stream 10.67% 和 3.97%。相比于 TSN＋TPN 与 GDN 网络,MTSC 的 Top－1 准确率分别提升了 4.27% 与 1.87%。同时,MTSC 超过了相同设置的 TSM 0.77%。针对 Top－5 准确率,以 8 帧作为输入的 TSM 和以 16 帧作为输入的 GDN 网络性能与以 8 帧作为输入的 MTSC 接近,但仍然低于 MTSC 0.07% 和 0.44%。同时,以 8 帧作为输入,MTSC 的 Top－5 准确率明显高于 TRN、TRN-2Stream 及 GDN。图 8 显示了部分类别的分类情况,对于 TSM 难以识别的“推某物使其旋转”类别,使用多时间尺度卷积取得了较大的提升,其他类别的识别数也获得了不同幅度的增加。这说明多时间尺度卷积使骨架网络获得了更强的时空特征提取能力。

3　结论

本文研究了 TSM,并利用公式推导了 TSM 可以等效为一组特殊的固定参数卷积核。同时,在分析过往基于卷积神经网络的行为识别模型的基础上,提出了多时间尺度卷积提取融合不同时间尺度的时空特征,以 ResNet50 为骨架构建了行为识别网络。

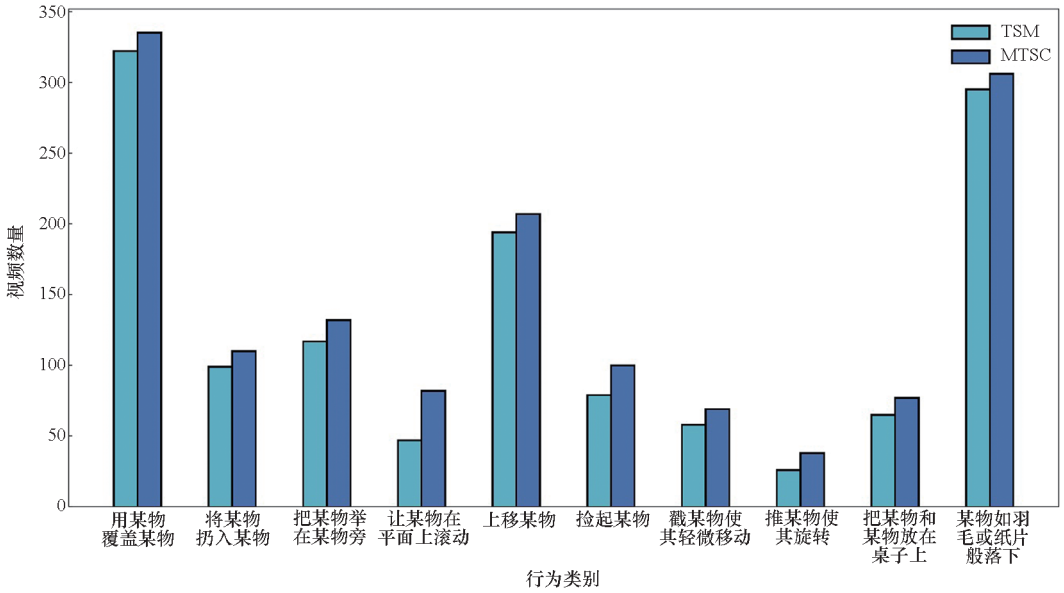

图8 TSM与多时间尺度卷积的部分类别识别情况

Fig. 8 Classification result of some behavior categories of TSM and MTSC

在行为识别网络构建方面,研究了多时间尺度卷积插入位置和控制截取特征通道数的参数 α 的取值对模型性能的影响。实验表明,当截取特征通道数为原通道数的1/4,插入位置为网络层1至网络层4时网络取得最好性能。通过实验对比验证了本文提出的网络优于TSM及其他网络,在Something-Something v2数据集上取得了59.47%的Top-1准确率。后续,将深入研究多时间尺度卷积瓶颈结构插入位置、如何选取截取特征以及降低网络参数量等问题,并更仔细地设计网络结构以取得更好的识别性能。

参考文献(References)

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2021-05-21]. https://arxiv.org/abs/1409.1556.

[3] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. (2014-11-12)[2021-05-21]. https://arxiv.org/abs/1406.2199.

[5] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Proceedings of the 14th European Conference on Computer Vision, 2016.

[6] ZHOU B L, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos[C]//Proceedings of the 15th European Conference on Computer Vision, 2018.

[7] ZOLFAGHARI M, SINGH K, BROX T. ECO: efficient convolutional network for online video understanding[C]//Proceedings of the 15th European Conference on Computer Vision, 2018.

[8] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2020.

[9] YANG C Y, XU Y H, SHI J P, et al. Temporal pyramid network for action recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[10] 刘董经典, 孟雪纯, 张紫欣, 等. 一种基于2D时空信息提取的行为识别算法[J]. 智能系统学报, 2020, 15(5): 900-909.
LIU D J D, MENG X C, ZHANG Z X, et al. A behavioral recognition algorithm based on 2D spatiotemporal information extraction[J]. CAAI Transactions on Intelligent Systems, 2020, 15(5): 900-909. (in Chinese)

[11] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2016.

[12] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with Pseudo-3D residual networks[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017.

[13] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[14] 张小俊, 李辰政, 孙凌宇, 等. 基于改进3D卷积神经网络的行为识别[J]. 计算机集成制造系统, 2019, 25(8): 2000-2006.

ZHANG X J, LI C Z, SUN L Y, et al. Behavior recognition method based on improved 3D convolutional neural network[J]. Computer Integrated Manufacturing Systems, 2019, 25(8): 2000 – 2006. (in Chinese)

[15] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? A new model and the kinetics dataset[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[16] XIE S N, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification[C]//Proceedings of the 15th European Conference on Computer Vision, 2018.

[17] QIU Z K, ZHAO X, HU Z L. Efficient temporal-spatial feature grouping for video action recognition[C]//Proceedings of IEEE International Conference on Image Processing (ICIP), 2020.

[18] 郭明祥, 宋全军, 徐湛楠, 等. 基于三维残差稠密网络的人体行为识别算法[J]. 计算机应用, 2019, 39(12): 3482 – 3489.
GUO M X, SONG Q J, XU Z N, et al. Human behavior recognition algorithm based on three-dimensional residual dense network[J]. Journal of Computer Applications, 2019, 39(12): 3482 – 3489. (in Chinese)

[19] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2020.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.