3D 多支路聚合轻量网络视频行为识别算法研究

胡正平[1,2],刁鹏成[1],张瑞雪[1],李淑芳[1],赵梦瑶[1]
(1. 燕山大学信息科学与工程学院,河北秦皇岛 066004;
2. 燕山大学河北省信息传输与信号处理重点实验室,河北秦皇岛 066004)

摘 要: 为构建拥有 2D 神经网络速度同时保持 3D 神经网络性能的视频行为识别模型,提出 3D 多支路聚合轻量网络行为识别算法. 首先,利用分组卷积将神经网络分割成多个支路;其次,为促进支路间信息流动,加入具有信息聚合功能的多路复用模块;最后,引入自适应注意力机制,对通道与时空信息进行重定向. 实验表明,本算法在 UCF101 数据集上的计算成本为 11.5GFlops,准确率为 96.2%;在 HMDB51 数据集上的计算成本为 11.5GFlops,准确率为 74.7%. 与其他行为识别算法相比,提高了视频识别网络的效率,体现出一定识别速度和准确率优势.

关键词: 深度学习; 神经网络; 行为识别
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2020)07-1261-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.07.003

Research on 3D Multi-Branch Aggregated Lightweight Network Video Action Recognition Algorithm

HU Zheng-ping[1,2], DIAO Peng-cheng[1], ZHANG Rui-xue[1], LI Shu-fang[1], ZHAO Meng-yao[1]
(1. *School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei* 066004, *China*;
2. *Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao, Hebei* 066004, *China*)

Abstract: To construct a video action recognition model with 2D neural network speed while maintaining the performance of 3D neural network, the 3D multi-branch aggregation lightweight network action recognition algorithm is proposed. Firstly, the neural network is divided into multiple branches by using grouped convolution. Secondly, to promote the information exchange between branches, a multiplexer module with information aggregation function is added. Finally, the adaptive attention mechanism is introduced to redirect channel and spatio-temporal information. Experiments show that, the computational cost of the algorithm on the UCF101 dataset is 11.5GFlops, and the accuracy is 96.2%; the computational cost on the HMDB51 dataset is 11.5GFlops, and the accuracy is 74.7%. Compared with other action recognition algorithms, it improves the efficiency of the video recognition network and reflects certain recognition speed and accuracy advantages.

Key words: deep learning; neural network; action recognition

1 引言

随着物联网时代视频信息的爆炸式增长,行为识别算法在安全监控,机器人设计,无人驾驶和智能家庭设计等方面都有广泛应用前景,因此逐渐成为国内外人工智能领域的一个研究热点问题[1~3].

近年来,3D 卷积神经网络(Convolutional Neural Network,CNN)利用其可对空间信息和时间信息同时建模的优势在行为识别领域取得了较大的进展[4~6],已成为视频理解的标准算法. 但是,3D CNN 的卷积核和特征图谱与 2D CNN 相比增加了额外的时间维度,导致 3D CNN 的参数量和计算量成指数级增长,难以满足实际应用中低延迟要求. 因此设计出低能耗且高精度的 3D CNN 就成为视频识别模型转化为实际应用的关键.

为解决 CNN 参数量和计算量过高的问题,研究思路主要包括:(1)减小卷积核尺寸;(2)降低特征通道数量或尺寸;(3)更新卷积计算方式;(4)降低网络复杂度. 从减小卷积核尺寸角度,文献[7]中使用多个 3×3

收稿日期:2019-09-06;修回日期:2020-03-17;责任编辑:孙瑶
基金项目:国家自然科学基金面上项目(No.61771420);河北省自然科学基金(No.F2016203422)

小卷积核的堆叠代替单个大卷积核,不但用更少的参数量保持感受野不变,而且更多的激活函数层也使网络获得更丰富的特征和更强的分辨能力. 从降低特征通道数量或尺寸角度,文献[8]中引入瓶颈结构,通过使用1×1卷积层缩减/扩张特征通道数,使3×3卷积层保持更小输入输出维度的瓶颈状态,从而减少参数量和计算量. 文献[8]中引入全局均值池化,通过减少全连接层输入节点数达到降低全连接层参数量目的. 文献[9]引入分组卷积操作,基本思路是将特征图谱分成n组进行卷积,每组的输入和输出通道数都降低为原来的$1/n$,各组之间相互独立,各组卷积完成后将输出叠加在一起作为该层输出,进而将参数量和计算量降低为原来的$1/n$. 其不足是单个组卷积的输入仅包含特征图谱中部分通道,且各组之间缺乏信息流通,将导致特征表示能力下降. 从更新卷积计算方式角度,文献[10]引入可分离卷积,将标准的卷积操作分为通道卷积和逐点卷积,先对每个输入通道用单个卷积核进行通道卷积,再使用1×1逐点卷积对通道卷积的输出进行线性组合,优点是资源占用情况有较大改善,缺点是性能下降较为明显. 从降低模型复杂度角度,文献[11]中引入网络剪枝操作,基本思路是仅保留权值大于阈值的重要连接,达到降低模型复杂度同时抑制过拟合的目的. 文献[12]提出基于K-means聚类的分块权值共享方案来减少参数量,进而减少点乘运算重复次数. 文献[13]使用知识蒸馏方法,将复杂、学习能力强的网络提取的"知识"蒸馏出来,传递给参数量小、学习能力弱的网络,以实现模型压缩的目的.

针对现有3D CNN计算开销较高的问题,从降低各支路特征通道数量和改进卷积机制角度,提出基于3D多支路聚合轻量网络的行为识别算法,以寻求计算效率和识别准确率间的平衡. 为降低计算开销,利用分组卷积将网络切分成独立卷积的多个支路;为改善支路间的信息交流,将多路复用器模块添加在每个残差块的头部,使并行支路间可共享特征信息;为进一步定位关键特征,使用3D自适应注意力模块对特征间的相关性进行建模以寻找核心信息位置. 本文设计了相关实验,以验证本算法能否在视频行为识别任务中降低资源开销,且保持较高的识别精度.

2 基于3D多支路聚合轻量网络的行为识别模型

2.1 模型整体结构

图1为基于3D多支路聚合轻量网络的行为识别模型的系统组成. 由于初始输入视频空间分辨率较大且冗余信息较多,故首先对输入视频片段进行下采样预处理,然后使用3D多支路聚合单元的堆叠结构进行特征提取,最后使用全局平均池化将输出特征图的时空分辨率降为1,再通过全连接层和Softmax层对视频进行分类.

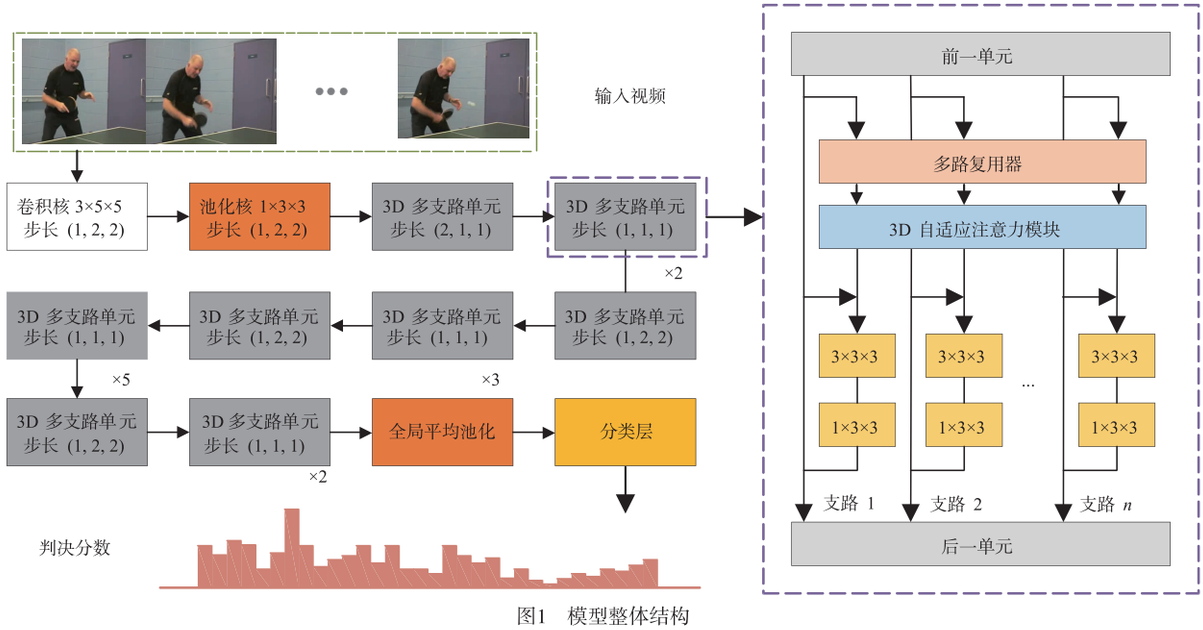

图1 模型整体结构

2.2 3D多支路聚合单元

如图2(c)所示,3D多支路聚合单元主要由残差切分单元、多路复用器和注意力模块三部分组成. 残差切分单元将主干网络沿通道维度分割成多个独立的平行支路;多路复用器使用两个1×1×1卷积层实现支路间信息共享的功能;注意力模块对各特征贡献程度进行

建模以确保每个支路有选择地关注关键特征.

2.2.1 残差切分单元

传统残差单元如图 2(*a*)所示,使用两个卷积层来学习特征,让 M_{in} 表示输入通道的数目,M_{mid} 表示中间通道的数目,M_{out} 表示输出通道的数目,假设卷积核尺寸为 $t\times k\times k$(在图 2 中 $t=k=3$),则传统残差单元的参数量为

$$\begin{aligned}para_b &= (t\times k\times k)\times M_{in}\times M_{mid}\\ &\quad +(t\times k\times k)\times M_{mid}\times M_{out}\\ &= (t\times k\times k)\times(M_{in}\times M_{mid}+M_{mid}\times M_{out})\end{aligned}\tag{1}$$

为减少总体计算成本,将特征图谱分割成 n 条平行且独立的路径,每个路径都与其他路径隔离,如图 2(*b*)所示,其参数量为

$$\begin{aligned}para_m &= ((t\times k\times k)\times(M_{in}/n\times M_{mid}/n\\ &\quad +M_{mid}/n\times M_{out}/n))\times n\\ &= \frac{1}{n}(t\times k\times k)\times(M_{in}\times M_{mid}+M_{mid}\times M_{out})\end{aligned}\tag{2}$$

式(1)和式(2)表明,残差切分单元在保持单元总通道数保持不变的前提下,参数量减少为传统残差单元的 1/n(本文中 n 值默认为 16,在 3.4.1 节将对该参数的选取进行详细分析). 考虑到 3D 卷积网络需要大量的训练数据防止过拟合且难以优化,将图 2(*b*)中第二个 3×3×3 卷积层替换为 1×3×3 卷积层以降低计算成本和优化难度. 值得注意的是:上述切片策略是基于残差单元特征图谱中跨通道相关性和通道内时空相关性的映射可以完全解耦的假设. 实际上,将各支路彼此隔离会阻碍它们间的任何信息交流,每条支路仅能访问少量的通道,从而限制整个模型的数据表示能力. 因此大多数采用分组卷积的网络仅对部分层进行切片,以防止丢失过多学习能力. 然而,非切分层的计算开销依旧未得到改善,成为降低能耗的主要瓶颈.

2.2.2 多路复用器

为恢复支路间的信息共享能力,在各残差支路头部附加轻量级瓶颈组件——多路复用器. 如图 2(*d*)所示,多路复用器使用两层 1×1×1 卷积层进行特征的收集和重分配功能,类似于 ResNet 中的瓶颈结构,第一个卷积层将通道数量降低为输出通道数的 1/c(本文中 c 设置为 4),第二个卷积层再将通道数恢复至输入通道数. 重复使用两层瓶颈 1×1×1 卷积层而非常规单层 1×1×1 卷积层的原因为前者在降低参数量的同时可多执行一次非线性激活使模型表达能力更强. 假设多路复用器的输入通道数为 i,输出通道数为 o(为保证处理前后特征维度不变,故输出通道数 o 等于输入通道数 i),常规单层 1×1×1 卷积层参数量为

$$para_p = (1\times 1\times 1)\times i\times o\tag{3}$$

两层瓶颈 1×1×1 卷积层参数量为

$$\begin{aligned}para_e &= (1\times 1\times 1)\times i\times\frac{o}{c}+(1\times 1\times 1)\times\frac{o}{c}\times i\\ &= \frac{2}{c}\times(1\times 1\times 1)\times i\times o\end{aligned}\tag{4}$$

由式(3)和式(4)可知,当 c 等于 4 时,后者参数量仅为前者的 1/2. 得益于多路复用器的信息交换功能,3D 多支路聚合单元可将整个网络按通道切片成多个平行且独立的分支而不受表达能力的限制.

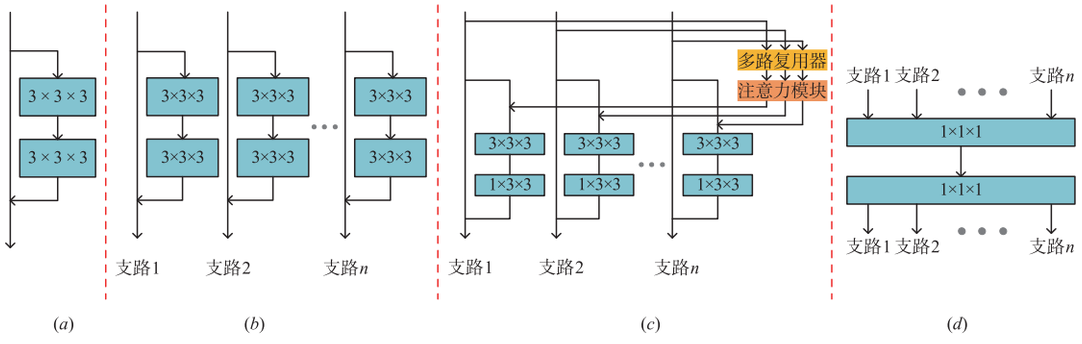

图2　从传统残差单元到3D多支路聚合单元

2.2.3 3D 自适应注意力模块

虽然多路复用器促进了支路间的信息交流,但是在多路复用器的所有特征中,很难区分开对某个支路有用的内容,这将降低信息交换的质量. 为解决多路复用器难以进行特征显著增强的问题,本文以 CBAM(Convolutional Block Attention Module,CBAM)[14] 为基础构建 3D-CBAM 注意力模块并嵌入多支路聚合单元中以确保每个分支有选择地更关注关键信息特征.

(1)注意力模块整体结构

3D 自适应注意力模块完整结构如图 3 所示,具体过程为:给定一个中间层特征图谱 $F\in R^{C\times T\times H\times W}$($C$ 表示通道数量,T 表示时序长度,H 和 W 表示空间分辨率)作为输入,先经过通道注意力模块生成 1D 通道注意力图谱 $M_c(F)\in R^{C\times 1\times 1\times 1}$($C$ 表示通道数量,1×1×1 表示时空分辨率)对输入数据 F 进行通道维度的特征重标定得到 F',再经过时空注意力模块生成 3D 时空注

意力图谱 $M_{TS} \in R^{1\times T\times H\times W}$（1 表示通道数量，$T\times H\times W$ 表示时空分辨率）对输入数据 F' 进行时空维度的特征重标定，得到最终输出 F''.

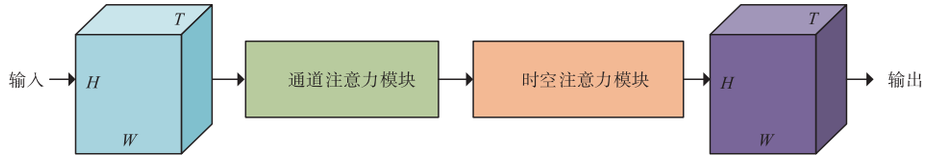

图3 3D-CBAM模块

(2)通道注意力模块

通道注意力模块基本思路是根据特征通道间的内部关系产生通道注意力映射，目的是关注重要特征并抑制次要特征，其结构如图4(*a*)所示.

特征图谱 $F\in R^{C\times T\times H\times W}$ 作为输入，首先，为充分聚集每个特征通道的时空信息，针对时空维度采用全局均值池化和最大值池化两种方式进行特征提取，得到两个 $F_C\in R^{C\times 1\times 1\times 1}$ 的通道描述. 再分别送入权值共享的多层感知机(Multi-Layer Perception，MLP)：第一层感知机输入神经元数为 C，输出神经元数为 $C/16$；第二层感知机输入神经元数为 $C/16$，输出神经元数为 C（在图4(*a*)中感知机输入神经元数用 In 表示，输出神经元数用 Out 表示）. 然后，将得到的两个特征相加后经过 Sigmoid 激活函数得到权重系数 $M_C(F)\in R^{C\times 1\times 1\times 1}$. 最后，将权重系数 $M_C(F)$ 和输入特征 F 相乘即可得到缩放后的新特征 $F'\in R^{C\times T\times H\times W}$，整个过程可描述为

$$\begin{aligned} M_C(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) \\ &\quad + MLP(\mathrm{MaxPool}(F))) \\ &= \sigma(W_1(W_0(F^c_{\mathrm{avg}})) + W_1(W_0(F^c_{\mathrm{max}}))) \end{aligned} \tag{5}$$

$$F' = M_C(F)\otimes F \tag{6}$$

其中 σ 表示 Sigmoid 函数，$W_0\in R^{C\times C/16}$ 是输入层到隐藏层的权重，$W_1\in R^{C/16\times C}$ 是隐藏层到输出层的权重. $\otimes$ 表示逐元素乘法，在乘法过程中，通道注意力图谱沿时空维度传播，时空注意力图谱沿通道维度传播.

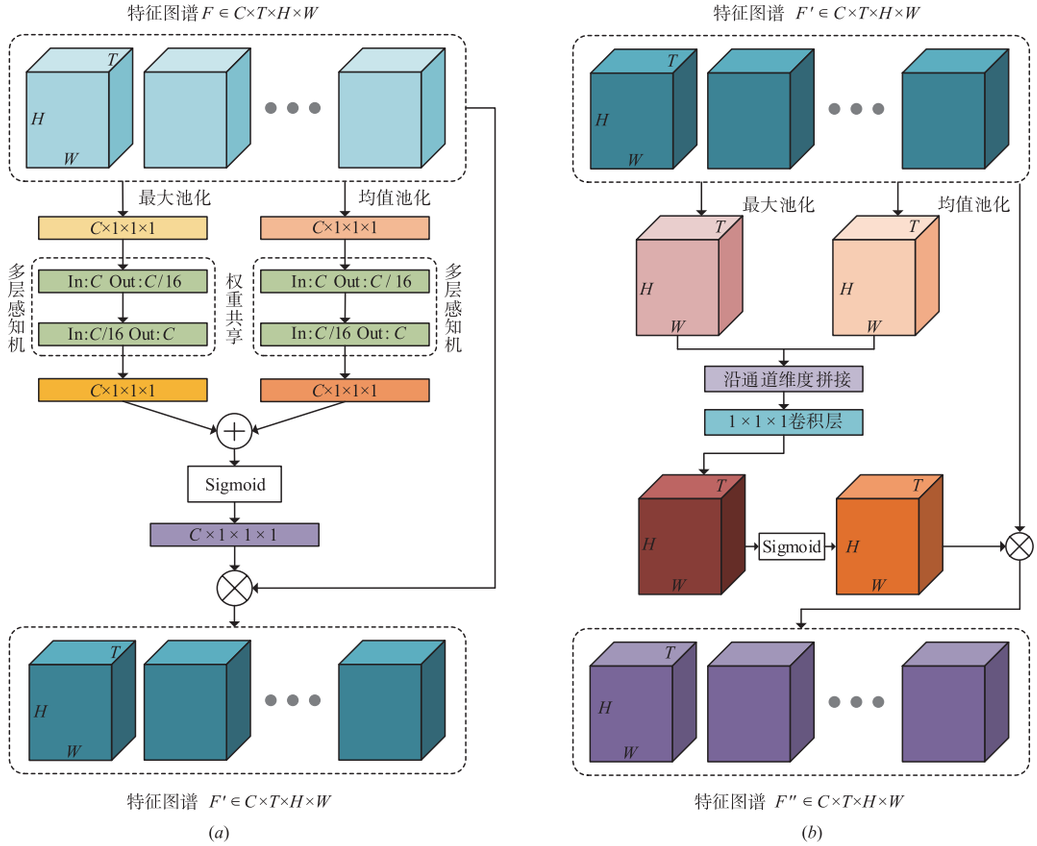

图4 通道注意力模块和时空注意力模块

(3)时空注意力模块

时空注意力模块和通道注意力模块间是相辅相成的,在执行通道注意力模块获取重点关注特征后,还需进一步获取其具体位置.

时空注意力模块的结构如图 4(*b*)所示,使用通道注意力模块输出的特征图谱 $F' \in R^{C\times T\times H\times W}$ 作为输入,首先,针对通道维度分别进行平均池化和最大池化得到两个 $F_{TS} \in R^{1\times T\times H\times W}$ 的通道描述,并将这两个描述沿着通道维度拼接在一起. 然后,经过一个卷积核为 7×7×7 的特征提取层和 Sigmoid 非线性激活层,得到权重系数 $M_{TS} \in R^{1\times T\times H\times W}$. 最后,将权重系数 M_{TS} 和输入特征 F' 相乘即可得到缩放后的新特征 $F'' \in R^{C\times T\times H\times W}$,整个过程可表示为

$$\begin{aligned} M_{TS}(F) &= \sigma(f^{7\times7\times7}[\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)])) \\ &= \sigma(f^{7\times7\times7}[F^{ts}_{\mathrm{avg}};F^{ts}_{\max}])) \end{aligned} \tag{7}$$

$$F'' = M_{TS}(F)\otimes F' \tag{8}$$

其中 σ 表示 Sigmoid 函数,$f^{7\times7\times7}$ 表示卷积核大小为 7×7×7 的卷积操作.

2.3　3D 多支路聚合网络

基于 2.2 节提出的 3D 多支路聚合单元构建 3D 多支路聚合网络(3D Multi-Branch Aggregation Network,3D MBA-Net),整个网络架构如表 1 所示. 在 Conv1 层,由于输入为彩色的 RGB 图像,通道数仅为 3,不足以进行分组操作,故在 Conv1 层依旧使用常规卷积. 在 Conv2 层与全局平均池化之间均使用 3D 多支路聚合单元构建网络,并仿照 ResNet-v2[15] 的设计,将归一化层和 ReLU 非线性激活函数放在卷积层之前进行预激活,以避免残差单元的输出恒为非负值导致特征表达能力下降. 在 Conv2 层的首个 3D 多支路聚合单元中使用步长为(2,1,1)的卷积层对输入特征信息执行时间维度上的下采样(16×56×56→8×56×56),在 Conv3、Conv4 和 Conv5 层的首个 3D 多支路聚合单元中依次使用步长为(1,2,2)的卷积层对输入特征信息执行空间维度上的下采样(8×56×56→8×28×28→8×14×14→8×7×7),在卷积层的最后使用全局平均池化将输出特征图的时空分辨率降为 1(8×7×7→1×1×1),再通过一个全连接层和 Softmax 层对预测结果进行打分.

3　实验与结果分析

本算法实验在 Kinetics、UCF101、HMDB51 三个数据集上进行,Kinetics 数据集包含来自 400 个人类活动类别的约 300000 个视频,每个视频持续 10 秒左右;UCF101 数据集包含来自如运动、乐器和人物交互等 101 个类别的 13320 个视频;HMDB51 数据集包含来自电影和网络视频等来源的 51 个类别的 6766 个视频.

表 1　3D 多支路聚合网络配置

层	单元重复次数	通道数	3D MBA-Net	
			输出尺寸	步长
Input		3	16×224×224	
Conv1	1	16	16×112×112	(1,2,2)
MaxPooling			16×56×56	(1,2,2)
Conv2	1	96	8×56×56	(2,1,1)
	2			(1,1,1)
Conv3	1	192	8×28×28	(1,2,2)
	3			(1,1,1)
Conv4	1	384	8×14×14	(1,2,2)
	5			(1,1,1)
Conv5	1	768	8×7×7	(1,2,2)
	2			(1,1,1)
AvgPooling			1×1×1	
FC			400	
参数量			8.4M	
计算量			11.5GFLOPs	

3.1　从零开始训练模型

为验证 3D MBA-Net 模型在从头训练运动特征时学习视频表示的有效性,使用大型 Kinetics 数据集进行训练. 考虑到 3D CNN 因参数量较大引起的收敛速度慢和难以优化的问题,使用和 I3D[16] 相同的参数初始方式:通过对 ImageNet-1K 数据集上预先训练过的 2D 模型中的参数进行时间维度上的“膨胀”来初始化 3D MBA-Net 模型. 然后,使用 3D MBA-Net 在 Kinetics 数据集上进行训练. 使用批处理随机梯度下降算法(批处理大小设为 1024),初始学习率设为 0.005,权重衰减设为 0.0001,动量设为 0.9,学习率衰减因子设为 0.1,在学习样本数为 1.8×10^7、2.8×10^7、3.4×10^7、3.8×10^7 时进行衰减.

表 2 给出 Kinetics 数据集上不同模型的对比结果,表中所有模型的输入仅使用 RGB 信息,不使用光流信息.

表 2　Kinetics 数据集上不同算法性能对比

模型	参数量/M	计算量/GFLOPs	Top-1 准确率/%	Top-5 准确率/%
Two-Stream[16]	12	–	62.2	–
ConNet + LSTM[16]	9	–	63.3	–
I3D-RGB[16]	12.1	107.9	71.1	89.3
R(2+1)D-RGB[17]	63.6	152.4	72.0	90.0
3DMBA-Net(Ours)	8.4	11.5	72.8	90.4

从结果中可以看出,基于 3D CNN 的模型比基于 2D CNN 的模型的 Top-1 精度更高,这一性能差距主要

是由于 2D CNN 仅能从单帧中提取特征,即使使用 LSTM 也难以从原始帧序列中建模复杂的运动特征,限制了其性能. 与 2D CNN 模型相比,这些 3D CNN 模型的参数量和计算成本较高. 相比之下,得益于高效的特征信息利用机制,3D MBA-Net 可有效减少时间维度带来的额外计算开销,使用低参数量和低计算量实现了高精度,表现出良好的计算效率和运动特征表示能力.

3.2 微调模型

在本节中,通过将 Kinetics 数据集上预训练的权重迁移到 UCF101 和 HMDB51 数据集进行微调,来评估 3D MBA-Net 的通用性和鲁棒性.

使用批处理随机梯度下降算法(批处理大小设为 128),初始学习率设为 0.005,权重衰减设为 0.0001,动量设为 0.9,学习率衰减因子设为 0.1,训练 UCF101 时在学习样本数为 2×10^5、4×10^5、6×10^5、7×10^5 时进行衰减,训练 HMDB51 时在学习样本数为 5×10^4、1×10^5、1.5×10^5 时进行衰减.

选取当下较为流行的 10 种行为识别算法进行对比实验,对比结果如表 3 所示.

表 3 UCF101 和 HMDB51 数据集上不同算法性能对比

模型	计算量/GFLOPs	是否使用光流	UCF101 准确率/%	HMDB51 准确率/%
ResNet-50[10]	3.8	否	82.3	48.9
ResNet-152[10]	11.3	否	83.4	46.7
CoViAR[18]	4.2	否	90.4	59.1
Two-Stream[19]	3.3	是	88.0	59.4
TSN[20]	3.8	是	94.2	69.4
C3D[21]	38.5	否	82.3	51.6
Res3D[22]	19.3	否	85.8	54.9
ARTNet[23]	25.7	否	94.3	70.9
I3D-RGB[16]	107.9	否	95.6	74.8
R(2+1)D-RGB[17]	152.4	否	96.8	74.5
3DMBA-Net(Ours)	11.5	否	96.2	74.7

从表 3 可以看出,在资源占用方面,3D MBA-Net 相比主流 3D CNN 计算开销大大减少,并大致与 2D CNN 处于同一级别,表现出良好的资源利用率;在识别性能方面,与主流行为识别算法相比,3D MBA-Net 虽未能达到目前最先进的水平,但在减少了大量的参数和计算开销的前提下,仍然保持了较为良好的识别性能.

3.3 模型运行速度对比

为验证本文提出的 3D MBA-Net 不仅能降低模型理论上的参数和计算成本,而且能提高模型的实际识别速度,故对主流的 3D 行为识别模型进行推理时间测试. 实验硬件为现阶段深度学习主流的 NVIDIA GTX 1080Ti 显卡,软件环境为 PyTorch 0.4,模型输入数据设置如下:批处理大小为 1,输入视频片段的长度和空间分辨率与原论文中保持一致,每帧图像数据为随机生成的 3 维数组以模仿 RGB 图像的 3 个通道,在输入至识别模型之前按照 ImageNet 数据集的均值和方差(mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])对其进行正则化处理,使其满足均值为 0,方差为 1 的数据分布. 为公平比较,排除了数据加载时间,并将所有模型的分类数统一设置为 100,每个模型均执行10^6 次测量以实现对不同数据集的通用性,并取其平均值作为最终结果并汇总在表 4 中.

表 4 模型运行速度对比

网络	输入数据尺寸	片段/秒	毫秒/片段
T3D[24]	32×224×224	14.41	69.41
I3D[16]	64×224×224	17.27	57.90
R(2+1)D[17]	16×112×112	23.25	43.01
S3D-G[25]	64×224×224	16.89	59.20
3DMBA-Net(Ours)	16×224×224	42.97	23.27

由表 4 结果可知,本文提出的 3D MBA-Net 在识别速度上相比 T3D、I3D、S3D-G 和 R(2+1)D 分别提高 198%、149%、84.8% 和 154%,低参数量和低计算开销的优势得以显现.

3.4 消融实验

在 UCF101 和 HMDB51 数据集上进行一系列消融实验,以证明提出的 3D 多支路聚合单元中各个组件的有效性.

3.4.1 切分支路数量对性能的影响

为评估选取不同分支数对算法性能的影响,将分支数量分别设置为 1、2、4、8、16 进行对比实验(分支数为 1 时等价于普通卷积),并将结果汇总在表 5 中.

表 5 不同分支数量对模型性能的影响

分支数量	UCF101			HMDB51		
	参数量/M	计算量/GFLOPs	准确率/%	参数量/M	计算量/GFLOPs	准确率/%
1	87.02	114.78	94.5	86.99	114.78	72.7
2	44.95	59.64	94.9	44.91	59.64	73.3
4	23.91	32.06	95.3	23.88	32.06	74.2
8	13.40	18.28	96.3	13.36	18.28	75.0
16	8.38	11.49	96.2	8.35	11.49	74.7

表 5 结果表明,在资源占用方面,分支数越大,模型的参数量和计算量越小,轻量级优势越明显,当分支数为 16 时,模型轻量化程度最高. 在识别准确率方面,分支数并不是越大越好:当分支数较少时,随着分支数的增加,准确率稳步上升;但当分支数较大时,随着分支数的增加准确率却会出现一定程度上的下降,当分支数为 8 时,模型准确率最高. 模型准确率随分支数增加而呈现先增后减趋势的原因是在各分支独立卷积过

程中,当分支数较少时,随着分支数的增加,各支路输入通道数也逐渐减少,在一定程度上可降低特征间的依赖性,迫使每个分支学习到更加独立的核心特征;但当分支数较多时,因输入通道数过少而导致各支路特征(通道)的组合方式甚至不能满足模型正常表述的最低要求,此时准确率就会逐渐下降.

综上所述,虽然分支数为 16 时准确率相比 8 时有所下降,但是准确率差距相较于计算节省完全处于可接受范围内,在识别准确率和资源占用之间取得了较好的均衡效果,因此将 $n=16$ 作为本文的默认参数.

3.4.2　多路复用器对准确率的影响

为测试多路复用器对模型准确率的影响,分别测试删除(表 6 中用 A 表示)/添加(表 6 中用 B 表示)多路复用器两种情况下的识别准确率.

表 6　删除/添加多路复用器对模型准确率的影响

数据集	A 方案准确率/%	B 方案准确率/%
UCF101	94.1	96.2
HMDB51	72.0	74.7

从表 6 可以看到,在 UCF101 和 HMDB51 数据集上,添加多路复用器的模型相比删除多路复用器的模型分别有 2.1% 和 2.7% 的性能提升,表明在支路之间共享信息的重要性.

3.4.3　注意力模块对准确率的影响

为验证注意力机制对行为识别准确率的影响,分别测试不使用任何注意力模块(表 7 中用 A 表示)、仅使用通道注意力模块(表 7 中用 B 表示)、仅使用时空注意力模块(表 7 中用 C 表示)和使用通道注意力模块+时空注意力模块(表 7 中用 D 表示)四种方案的识别准确率,结果如表 7 所示.

表 7　注意力机制对模型准确率的影响

数据集	A 方案准确率/%	B 方案准确率/%	C 方案准确率/%	D 方案准确率/%
UCF101	95.1	96.0	95.7	96.2
HMDB51	73.4	74.2	73.9	74.7

由表 7 可知,识别效果为:D 方案 > B 方案 > C 方案 > A 方案.一方面,B、C、D 方案的准确率都比不含注意力机制的 A 方案更高,显示出注意力模块对于准确率的提升.另一方面,B 方案 > C 方案的原因是两模块具有不同的作用域:从时空角度看,通道注意力应用于全局,而时空注意力应用于局部.为追求识别率,可将两者结合使用(D 方案),在通道注意力选定与动作类别紧密关联的目标区域后使用时空注意力在此基础上进行更精准的定位.

4　结论

为解决 3D 卷积计算开销过高导致难以实际应用的问题,本文提出一种 3D 多支路聚合网络,通过将特征通道分解为并行且独立的支路,达到降低参数数量和计算开销的目的,添加多路复用器并使用自适应注意力机制对支路间特征信息进行共享和重新校准,获取更高的识别精度.实验结果表明,与当前先进的识别算法相比,本文算法有效降低了 3D 卷积神经网络的参数数量和计算开销,兼顾了识别速度和识别精度.

参考文献

[1] 罗会兰,王婵娟. 行为识别中一种基于融合特征的改进 VLAD 编码方法[J]. 电子学报,2019,47(1):49 – 58.
LUO Hui-lan,WANG Chan-juan. An improved VLAD coding method based on fusion feature in action recognition [J]. Acta Electronica Sinica,2019,47(1):49 – 58. (in Chinese)

[2] 张友梅,常发亮,刘洪彬. 基于 3D 人体骨架的动作识别[J]. 电子学报,2017,45(4):906 – 911.
ZHANG You-mei, CHANG Fa-liang, LIU Hong-bin. Action recognition based on 3D skeleton[J]. Acta Electronica Sinica,2017,45(4):906 – 911. (in Chinese)

[3] 罗会兰,童康,孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报,2019,47(5):1162 – 1173.
LUO Hui-lan, TONG Kang, KONG Fan-sheng. The progress of human action recognition in videos based on deep learning:a view[J]. Acta Electronica Sinica,2019,47(5):1162 – 1173. (in Chinese)

[4] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. USA:IEEE,2017. 5533 – 5541.

[5] Xu H, Das A, Saenko K. R-c3d:Region convolutional 3d network for temporal activity detection[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. USA:IEEE,2017. 5783 – 5792.

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

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [A]. Computer Science[C]. ICLR Press,2014. 1549 – 1556.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE,2016. 770 – 778.

[9] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 1492 – 1500.

[10] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 4510 – 4520.
[11] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[A]. Advances in Neural Information Processing Systems[C]. Canada: Curran Associates, 2015. 1135 – 1143.
[12] Wu J, Leng C, Wang Y, et al. Quantized convolutional neural networks for mobile devices[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 4820 – 4828.
[13] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38 – 39.
[14] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[A]. Proceedings of the European Conference on Computer Vision[C]. Germany: Springer, 2018. 3 – 19.
[15] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[A]. European Conference on Computer Vision[C]. The Netherlands: Springer, 2016. 630 – 645.
[16] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2017. 6299 – 6308.
[17] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 6450 – 6459.
[18] Wu C Y, Zaheer M, Hu H, et al. Compressed video action recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 6026 – 6035.
[19] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[A]. Advances in Neural Information Processing Systems[C]. Canada: Curran Associates, 2014. 568 – 576.
[20] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[A]. European Conference on Computer Vision[C]. The Netherlands: Springer, 2016. 20 – 36.
[21] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. USA: IEEE, 2015. 4489 – 4497.
[22] Tran D, Ray J, Shou Z, et al. Convnet architecture search for spatiotemporal feature learning[A]. Computer Science[C]. ICLR Press, 2017. 1708 – 1716.
[23] Wang L, Li W, Li W, et al. Appearance-and-relation networks for video classification[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 1430 – 1439.
[24] Diba A, Fayyaz M, Sharma V, et al. Temporal 3D ConvNets using temporal transition layer[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops[C]. USA: IEEE, 2018. 1117 – 1121.
[25] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[A]. Proceedings of the European Conference on Computer Vision[C]. Germany: Springer, 2018. 305 – 321.

作者简介

胡正平(通讯作者) 男,1970 年 8 月出生,四川仪陇人. 燕山大学信息科学与工程学院教授,博士生导师,主要研究方向为模式识别、图像处理.
E-mail: hzp@ysu.edu.cn

刁鹏成 男,1996 年 1 月出生,河北衡水人. 燕山大学信息科学与工程学院硕士研究生,主要研究方向为行为识别.
E-mail: ysdpc666@sina.com

张瑞雪 女,1994 年 12 月出生,黑龙江齐齐哈尔人. 燕山大学信息科学与工程学院硕士研究生,主要研究方向为视频分类.

李淑芳 女,1981 年 5 月出生,河北滦南人. 燕山大学信息科学与工程学院博士研究生,主要研究方向为模式识别.

赵梦瑶 女,1995 年 10 月出生,黑龙江牡丹江人. 燕山大学信息科学与工程学院博士研究生,主要研究方向为视频异常检测.