文章编号:1671-4598(2022)03-0251-08 DOI:10.16526/j.cnki.11—4762/tp.2022.03.042 中图分类号:TP391 文献标识码:A

基于深度残差网络的人体行为识别算法研究

冯 宇,席志红
(哈尔滨工程大学 信息与通信工程学院,哈尔滨 150001)

摘要:针对原始C3D卷积神经网络的层数较少、参数量较大和难以关注关键帧而导致的人体行为识别准确率较低的问题,提出一种基于改进型C3D的注意力残差网络模型;首先,增加原始网络卷积层并采用卷积核合并与拆分操作实现(3×1×7)和(3×7×1)的非对称式卷积核,之后采用全预激活式残差网络结构来增加构建的非对称卷积层,并且在残差块中增加时空通道注意力模块;最后,为展示该算法的先进性和应用性,则将该算法与原始C3D网络以及其他流行算法分别在基准数据集HMDB51和自建的43类别体育运动数据集上相比较;实验结果表明,该算法与原始C3D网络相比,在HMDB51和43类体育运动数据集上分别提高了9.88%和21.61%,参数量比原来降低了38.68%,并且结果也优于其他流行算法。

关键词:深度学习;三维卷积;非对称式卷积核;残差网络;注意力模块;人体行为识别

Research on Human Action Recognition Algorithm Based on Deep Residual Network

FENG Yu, XI Zhihong
(College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: Aiming at the problem that the original C3D convolutional neural network has a small number of layers, a large amount of parameters, and the difficulty of focusing on key frames lead to the low accuracy of human behavior recognition, an improved C3D-based attention residual network model is proposed. First, adds the original network convolution layer and uses the convolution kernel merge and splits operation to realize the asymmetric convolution kernel of (3×1×7) and (3×7×1), and then the fully pre-activated residual network structure is used to increase the constructed asymmetric convolutional layer, and the spatiotemporal channel attention module is added to the residual block. Finally, in order to demonstrate the advancement and applicability of the algorithm, the algorithm is compared with the original C3D network and other popular algorithms on the benchmark data set HMDB51 and the self-built 43 categories sports data set. Experimental results show that compared with the original C3D network, the algorithm has increased by 9.88% and 21.61% on the HMDB51 and 43 types of sports data sets, respectively, and the quantity of parameters has been reduced by 38.68%, and the results of the algorithm are better than that of other popular algorithms.

Keywords: deep learning; three-dimensional convolution; asymmetric convolution kernel; residual network; attention module; human behavior recognition

0 引言

近年来随着对于计算机视觉领域研究的不断深入,越来越多的技术成果在不断满足人们生产生活的需求,人体行为识别技术也随之受到越来越多的关注,应用场景也变得越来越丰富,例如虚拟现实技术、视频监控领域和医疗健康等方面,目前研究人体行为识别技术的方法主要有基于深度学习的方法和传统的基于手工提取特征的方法。

基于手工提取特征的方法实际上就是对特定的视频图像采用传统的机器学习算法先提取其中的人体行为目标局部或者全局特征,然后对提取的特征采取编码以及规范化的形式,最后通过训练构建好的模型来得到预测分类结果。目前在传统方法中采用局部特征提取的方法应用较为广泛,其中Laptev等[1]通过将Harris特征角点检测方法扩展到三维时空中,提出了时空兴趣点(STIPs, spatio- temporal interest points)。之后通过不断对时空特征的研究发现可以将STIPs与方向梯度直方图(HOG, histogram of oriented gradient)[2]等局部描述子结合,采用聚类降维以及词袋模型和分类器相结合的方法进行姿态识别。Richardson等[3]提出了马尔科夫逻辑网络(Markov Logic Networks),该网络对动作之间的时空关系进行描述,改善了复杂人体姿态情况的识别效果。Wang等采用光流轨迹对视频帧间的时序关系进行模仿,提出密集轨迹(DT, dense trajectory)算法[4]用于人体行为识别,为去除由于相机运动而对特征提取造成的影响,则对光流图像进行优化,进一步提出改进的密

收稿日期:2022-01-18; 修回日期:2022-01-27。
基金项目:国家自然科学基金资助项目(60875025)。
作者简介:冯 宇(1997-),男,黑龙江尚志人,硕士研究生,主要从事深度学习、视频人体行为识别方向的研究。
席志红(1965-),女,黑龙江哈尔滨人,教授,博士,硕士研究生导师,主要从事图像处理与应用方向的研究。
引用格式:冯 宇,席志红.基于深度残差网络的人体行为识别算法研究[J].计算机测量与控制,2022,30(3):251-258.

集轨迹（iDT，improved dense trajectory）算法[5]用于人体行为识别。虽然基于手工提取特征的方法相对较为成功，但是该类方法是针对固定视频设计提取特征，无法满足输入视频的通用性，并且计算速度非常慢，很难满足现实世界中实时性的要求。

基于深度学习的行为识别方法主要是先通过设计好的神经网络对输入视频进行自动行为特征提取并不断训练模型，之后将训练好的模型用于分类识别。目前基于深度学习的行为识别算法主要通过三维卷积神经网络、双流网络、循环卷积神经网络和注意力网络进行构建。早在 2013 年 Ji 等[6]通过详细的实验提出了使用三维卷积神经网络提取视频的时空特征来对人体行为进行有效识别。随后 Simonyan 等[7]利用视频的图像帧和光流帧分别输入空间流网络和时间流网络来提取时空信息，提出用于行为识别的双流卷积架构。Tran 等[8]通过系统的实验研究找到了最适合行为识别的三维卷积核尺寸，并提出 C3D 网络用于直接提取时空特征进行行为识别。Donahue 等[9]利用循环卷积神经网络（RNN）能够针对时间序列很好建模的优势，提出长时循环卷积网络（LRCNN）用于视频的行为识别。Li 等[10]提出带有视觉注意力机制的深度视觉注意模型用于行为识别。Liu 等[11]通过将视频帧内空间信息的稀疏性引入到空间维度上，提出一种轻量级的组帧网络（GFNet）用于行为识别。Yang 等[12]提出一种即插即用的特征级时间金字塔网络（TPN）用于行为识别。

虽然基于深度学习的人体行为识别方法较手工提取特征的方法快捷、方便，并且时效性好，但是在实际应用中也很容易受到光线、背景杂乱、摄像视角等复杂环境因素影响，所以还需要在模型抗干扰方面继续优化深度神经网络模型。

本文基于 C3D 网络[8]，提出一种基于 C3D 注意力残差网络模型用于人体行为识别。本文算法在 C3D 原始网络基础上通过三维卷积核的合并与拆分以及全局平均池化来大量减少网络参数，达到压缩网络的效果，并且使用软池化（SoftPool）代替原有的最大池化（Maxpool）操作，以最大程度减小池化的信息损失，之后采用分组归一化（GN，group normalization）对网络进行正则化处理，为进一步提取深度特征采用全预激活形式的残差结构来增加三维卷积层，最后为能更好的关注视频的关键帧，在网络中引入时空通道注意力模型来提高网络模型的识别能力和抗干扰性能。

1 C3D 神经网络

在本文中提出一种基于 C3D 注意力残差网络的行为识别模型来改善原始 C3D 网络的不足。将在本节中简要介绍原始 C3D 网络，并在下一节介绍基于残差网络和注意力网络的人体行为识别算法。

C3D 神经网络是非常经典的通过三维卷积直接提取时空特征并用于视频人体行为识别的卷积神经网络。该三维卷积网络证明了比二维卷积网络更适合时空特征的学习，不需要复杂的提取视频光流帧图象。该网络还通过采用多种卷积核尺寸大小进行实验，验证了采用大小为（3×3×3）的卷积核性能最好。C3D 网络是一个总共具有 10 层的深度卷积神经网络，其中具有 8 层的三维卷积层，2 层的全连接层。该网络是以图像尺寸调整为（112×112）的 3 通道 16 帧视频图像作为输入，并且网络中为了能够更好的提取特征，则将特征提取与池化功能分开，即网络中所有三维卷积层只进行提取时空特征，使得经过卷积层的输入与输出尺寸相同，池化功能则全部采用三维最大池化层来完成。为了使得时间信息不被过早地丢失，因此仅第 1 个三维池化内核的尺寸为（1×2×2），而后续的三维池化内核均为（2×2×2）。最后将提取到的特征输入到后面的 2 层全连接层中进行特征分类，并最终通过 Softmax 分类器将视频中各个人体动作类别的分类概率进行输出。C3D 神经网络的整体结构如图 1 所示。

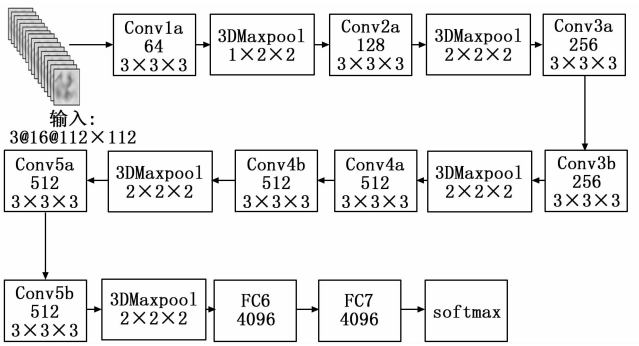

图 1 C3D 网络结构

2 人体行为识别算法

2.1 基于残差网络的人体行为识别算法

残差网络最早由 He 等[13]为解决更深层网络更难训练的问题而提出的一种能够优化网络训练的结构，其中的跳转连接也正是该结构能够优化训练网络的关键。由于残差网络的诞生，避免了为提取深度特征而增加卷积层所引起的网络退化问题，同时也进一步发展了人体行为识别领域。Tran 等[14]通过将三维卷积核应用到二维残差网络中提出了新型的 Res3D 网络架构。Qiu 等[15]采用伪三维块（P3D）替代原始二维残差网络的二维残差单元，提出一种伪三维残差网络（P3D ResNet）。Tran 等[16]通过将时空卷积残差块分解为空间卷积块和时间卷积块，提出一种残差块为（2+1）D 卷积块的新型残差网络。

2.2 基于注意力的人体行为识别算法

卷积神经网络在提取特征时会将视频中的每一帧图像和图像中的每一个像素都视为同等重要，这将导致计算资源的浪费以及针对特定任务的识别性能下降等问题。Sharma 等[17]通过利用在时间和空间都具有深层序列的长短时记忆（LSTM）单元构建深层递归神经网络（RNNs）来提取时空特征，并结合带有双重随机惩罚项和注意力正则化项

的交叉熵损失函数，提出一种基于软注意力的视频人体行为识别模型。Liu 等[18]采用带有 LSTM 单元的递归神经网络（RNNs）来不断地调整特征注意力权值，从而提出时间注意力模型，该模型能够对输入的特征序列进行时域扫描，然后再利用 RNN 的不断迭代来判别当前帧的重要性，并对所关注的特征进行加权，以此通过选择性地关注视频帧进行高效人体行为识别。Dai 等[19]通过构建时域特征流和时空特征流的 LSTM 双流网络来进行人体行为识别，其中时域特征流是将光流图像输入到时间注意力模块中来自动确定每个光流图像中的主要区域，并将这些关键信息图像的特征表示进行聚合，从而以特征向量形式来呈现光流图像的主要特征，时空特征流则是在池化层之后建立 LSTM 网络，用以学习空间深度特征之间内在的时域关系，并引入时空注意力模块来对不同层次的深层特征赋予不同程度的权重。Zhao 等[20]采用基于光流分析的自适应关键帧提取策略将视频中的关键帧预先提取，之后将提取好的视频关键帧作为输入序列，利用 C3D 神经网络对该关键帧序列进行特征提取，最后将提取好的视频关键帧特征输入到训练好的支持向量机中进行人体行为识别。

3　实现方法

3.1　非对称式三维卷积层

由于原始 C3D 卷积网络参数量较大，所以为减少网络参数量并同时增加特征提取能力，使整体网络结构轻量化，则本文采用卷积核的合并与拆分操作实现能够大量减少网络参数以及加强特征提取效果的非对称式三维卷积层，并且将全连接层使用全局平均池化（GAP，Global Average Pooling）操作进行替换。Inception- v3 网络[21]中提到可以将任何一个（$n\times n$）卷积采用（$1\times n$）和（$n\times 1$）卷积进行替代，n 越大这种非对称分解方式越能体现资源的节省，并且还提到本质上可以采用两层（3×3）卷积替换一层（5×5）卷积。

基于以上思想将其扩展到时空领域，本文先增加三维卷积层（3×3×3），使原始 C3D 网络除开始的第一个三维卷积层外每一个卷积层部分都是 3 个具有卷积核为（3×3×3）的三维卷积层。之后将这 3 个卷积层合并成卷积核为（3×7×7）的一个三维卷积层。这样变换可以比原始 C3D 网络具有更大的空间感受野，增强了特征提取能力，并且使得带有卷积核为（3×7×7）的一个三维卷积层就与卷积核为（3×3×3）的 3 个三维卷积层在空间域上具有相同的感受野，同时还起到了大量减少网络参数以及计算量的效果，其在空间域上的合并等效原理如图 2 所示。为进一步减少网络参数、增加空间特征的多样性以及加速网络训练，将带有卷积核为（3×7×7）的三维卷积层进行非对称式拆分成卷积核为（3×1×7）和（3×7×1）的两个三维卷积层，即得到两个非对称式三维卷积层，非对称式拆分原理如图 3 所示。

全连接层往往是网络参数量最大的层，并且往往因巨

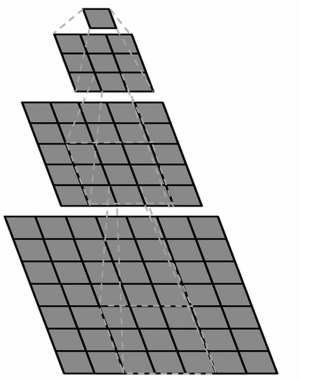

图 2　7×7 合并等效原理图

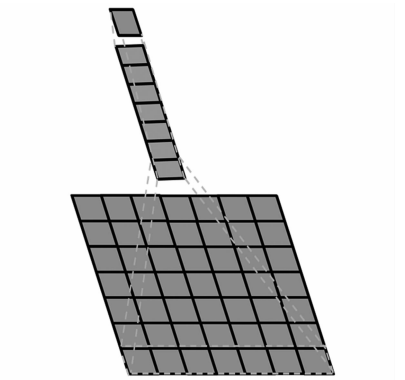

图 3　7×7 非对称式拆分原理图

大的网络参数量非常容易引起网络的过拟合现象，所以一种可以将全连接层替代的全局平均池化[22]方法被提出来，如图 4 所示。由于全局平均池化操作是对输入特征直接池化而不包含神经元，所以可以节省大量网络参数。由全局平均池化操作输出的图像尺寸为（1×1×1），所以可以避免由于输入网络原始图像尺寸的不同而后期造成网络维度不匹配的问题。

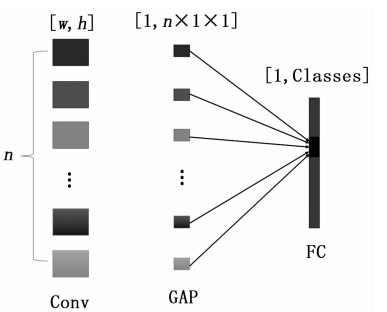

图 4　全局平均池化示意图

3.2　引入全预激活式残差结构

获取更深层次的网络特征往往需要加深网络，但网络层数的增加很容易造成过拟合和网络退化现象，所以为避免以上问题本文采用了残差式网络结构。

本文没有采用传统形式的残差网络结构[13]，而是采用全预激活式残差网络结构[23]并将其扩展到时空领域。因为

原始形式的残差连接在网络干路上存在激活函数，结构如图 5 所示，其只是在残差块中形成恒等映射，而没有在由残差块组成的网络中形成真正的恒等映射，所以会阻碍信息的传递，可能还会导致网络最终达不到最优化结果。

全预激活式残差连接由正则化和激活函数组成信息进入卷积权重层前的预激活操作，结构如图 6 所示。这种预激活操作不仅可以优化网络最终结果，还可以对网络模型起到正则化效果。全预激活式残差网络结构在整个网络中可以形成一个直接通路，信息可以在任意两个残差块之间直接传递，这样更有利于信息的流通和加速网络训练，并且这种网络结构还可以对所有将要进入卷积权重层的输入进行正则化，有利于改善模型最终识别结果。

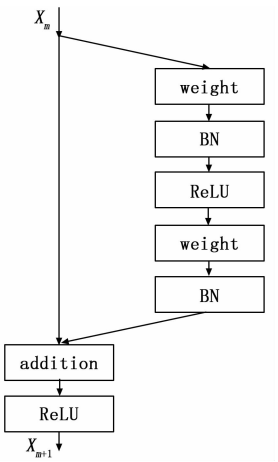

图 5 原始残差结构图

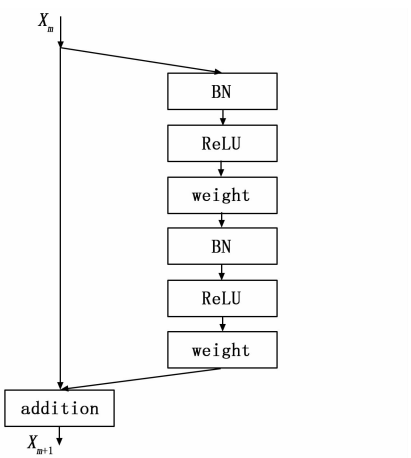

图 6 全预激活式残差结构图

3.3 软池化替代最大池化操作

在卷积神经网络中为了增加感受野和减少卷积过程中的计算需求，所以采用池化操作减小激活特征图的尺寸。以往的网络模型常采用最大池化或平均池化，但是最大池化是对池化区域内取最大激活值，这样非常容易造成大量的信息损失，属于一种暴力池化，平均池化则是对池化区域的所有激活值取平均，这样会降低池化区域中所有激活值对特征图的影响，属于一种抑制性池化。为了能够平衡最大池化和平均池化的消极影响，同时利用两者的池化优势，所以本文网络采用介于两者之间的软池化（SoftPool）[24]来替代原始 C3D 网络中的最大池化。

SoftPool 是一种快速有效的基于指数加权和的池化方法，该方法首先通过基于 Softmax 指数归一化的方式得到应用于激活特征图的每个激活值的权重，之后通过对池化核区域内的每个激活值进行权重加权求和来得到最终的 SoftPool 输出。该权重定义如式（1）所示：

$$W_i = \frac{\exp(a_i)}{\sum_{j\in R}\exp(a_j)} \tag{1}$$

式中，R 为池化核区域，a_i 和 a_j 表示为激活特征图池化核内的激活值，W_i 为分配给池化核内每个激活值的权重，i 和 j 表示为池化核区域的索引号。最终的 SoftPool 输出如式（2）所示：

$$\tilde{a} = \sum_{i\in R} W_i * a_i \tag{2}$$

式中，W_i 为权重，a_i 为激活特征图池化核内的激活值，R 为池化核区域，$\tilde{a}$ 为最终软池化输出结果，i 为池化核区域的索引号。这种分配指数权重的方式可以确保较大的激活值能够对输出产生更大的影响，而较小的权重也能够展现自己的贡献而不至于被完全抑制。该方法在很大程度上保留了原有的特征属性，同时放大了更高强度的特征激活，并且不需要训练参数，所以这种池化方式更有助于提高分类识别精度以及加速网络训练。

3.4 分组归一化

为改善网络训练过程中输入数据的分布情况，使得各三维卷积层接收的输入数据分布一致以及减轻过拟合现象发生，本文引入分组归一化（GN，group normalization）[25]对网络进行正则化操作。

目前大多数网络使用的正则化操作有批归一化（BN，batch normalization）[26]、Dropout[27]等。但是 Dropout 正则化大多数用于全连接层后面，本网络由于去掉了全连接层所以没有采用 Dropout 正则化。BN 正则化通过在 Batch 内计算输入数据的均值和方差进行归一化特征操作，虽然该操作可以起到很好的网络正则化效果，并且能够简化深层网络的优化，但是该方法却严重依赖 Batch 的大小，Batch 大小的不同不仅会严重影响最终分类识别的结果，还会对内存的占用产生巨大影响，并且会导致训练好的网络模型难以迁移到小型设备中。

基于以上，本文采用 GN 操作对各三维卷积层进行正则化，GN 正则化方法是通过将通道分组，然后在分好的组内计算用于归一化特征的均值和方差，原理如图 7 所示。假设要归一化的输入数据为 $x=[x_1,x_2,\cdots,x_d]$，那么其中的第 k 个输入 x_k 的分组归一化结果 y_k 如式（3）所示：

$$y_k = \gamma\frac{x_k-\mu}{\sqrt{\delta^2+\varepsilon}}+\beta \tag{3}$$

式中，μ 和 δ^2 分别是输入 x 的均值和方差，ε 是一个防止分母为零的小量，γ 和 β 是 GN 在分组中每个通道的可学习参数

用以增强网络表达能力。其中用来计算 x 的均值和方差的像素集合 S_i 在 GN 中定义如式（4）所示：

$$S_i = \left\{ k \mid k_N = i_N, \left| \frac{k_C}{\frac{C}{G}} \right| = \left| \frac{i_C}{\frac{C}{G}} \right| \right\} \tag{4}$$

式中，G 为分组数，C 为需要分组的通道数，N 表示批量，$|\cdot|$ 表示向下取整运算，式子 S_i 说明 GN 是以数目为 $\frac{C}{G}$ 的通道数沿着（H,W）轴，即沿着空间高度和空间宽度轴来计算输入 x 的均值和方差。由该正则化方法的计算方式可知，GN 正则化与 Batch 大小无关，摆脱了内存消耗的限制，并且 GN 被论文[25]证实在多种计算机视觉任务中其性能均优于 BN 和其他归一化方法。

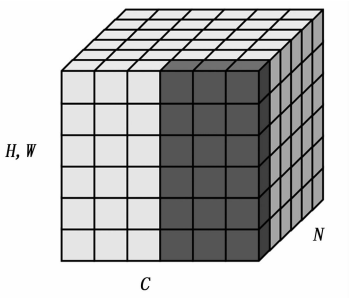

图 7　GN 原理图

3.5　时空通道注意力模块

人体行为识别需要将视频片段首先处理成时间序列的视频帧，然后再送入网络进行分类识别，然而一个视频片段中能够精准识别动作的关键帧往往包含在大量冗余帧中，所以网络中需要能够产生关键帧信息的注意力模块。

本文采用基于卷积块注意力模型（CBAM）[28]改进的改进型卷积块注意力模型（iCBAM）[29]来产生用于精准分类识别的注意力特征图。CBAM 能够沿着通道和空间两个不同维度产生最终的注意力特征图，而 iCBAM 在其基础上加入了时间维度，将其扩展到了时空领域，成为能够对通道、空间、时间 3 个方面进行充分关注的注意力模块。对于人体行为识别任务来说通道注意力集中在对给定的输入图像需要关注的是“什么”，空间注意力则集中在“哪里”是信息丰富的部分，时间注意力则是找到“哪些”是关键帧。iCBAM 会依次沿着通道、空间、时间 3 个维度来产生注意力特征图，并在这个过程中会将通过每一个维度而输出的特征与该维度的输入特征相乘来进行自适应的特征细化以产生最终的注意力特征图，原理如图 8 所示。

首先获得一个经过三维卷积提取的特征图为 $F \in \mathbf{R}^{C\times M\times H\times W}$ 作为 iCBAM 的输入，式中 R 为网络中的时空域，C 为网络通道数，M 为视频帧数，H 为视频帧图象的高度，W 为视频帧图像的宽度。该特征图先通过通道注意力模块，获得 1D 通道注意力特征图即 $T_C(F)$，其中 $T_C \in \mathbf{R}^{C\times 1\times 1\times 1}$ 为通道注意力模块，之后与原始特征图 F 逐元素相乘得到经过自适应特征细化的通道注意力特征图 F' 如式（5）所示：

$$F' = T_C(F) \otimes F \tag{5}$$

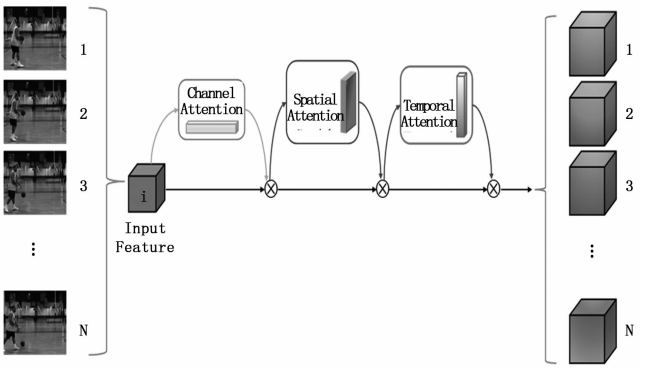

图 8　iCBAM 结构图

式中，$\otimes$ 为逐元素相乘。该新得到的特征图再通过空间注意力模块，获得 2D 空间注意力特征图即 $T_S(F')$，其中 $T_S \in \mathbf{R}^{1\times 1\times H\times W}$ 为空间注意力模块，随后再与特征图 F' 逐元素相乘得到新的自适应特征细化后的特征图 F'' 如式（6）所示：

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

经过特征细化后的特征图为进一步从视频帧中找出关键帧，所以再通过一个时间注意力模块即 $T_M \in \mathbf{R}^{1\times M\times 1\times 1}$ 来区别出关键视频帧，最终原始特征图 F 经过自适应特征细化后得到的时空通道注意力特征图 $F^{''}$ 如式（7）所示：

$$F^{''} = T_M(F'') \otimes F'' \tag{7}$$

基于 C3D 注意力残差网络的人体行为识别算法整体结构如图 9 所示。

4　实验

4.1　实验环境和相关设置

本文实验采用的硬件设备配置为 Inter Core i7- 8700 CPU，3.2 GHz，16 GB RAM，1T SSD，Nvidia Tesla T4 (16 GB) GPU，平台操作系统为 Ubuntu16.04，编程语言为 Python3.7，深度学习框架采用 PyTorch1.6.0 版本。

网络中的迭代周期（Epoch）设为 50 次，初始学习率设为 0.000 1，并且每经过 10 次迭代周期后将以 0.1 进行衰减，训练采用的批量大小为 8，分组归一化中采用的分组数为 32，实验采用 GELU 激活函数[30]，并且使用 Adam 优化算法来优化网络。

4.2　视频行为识别数据集和评价标准

本文使用人体行为识别公共基准数据集 HMDB51 和自建的 43 类别体育运动数据集进行实验。HMDB51 数据集共有 6 766 段视频剪辑，包含 51 个人体行为类别，并且每个类别的视频数量都不少于 101 个，帧率为 30 fps，视频的分辨率为 320×240，该数据集动作主要可分为 5 类，分别为常见的单独面部动作如微笑等、操纵物体并伴有面部动作如喝水等、一般身体动作如跳水等、与物体互动的身体动作如骑车等、人际互动的身体动作如握手等。该数据集中的视频大部分涉及摄像机的抖动、遮挡、不同拍摄角度以及低质量帧的问题，所以该视频数据集在人体行为识别任务中具有一定挑战性。由于 HMDB51 数据集广泛包含人体

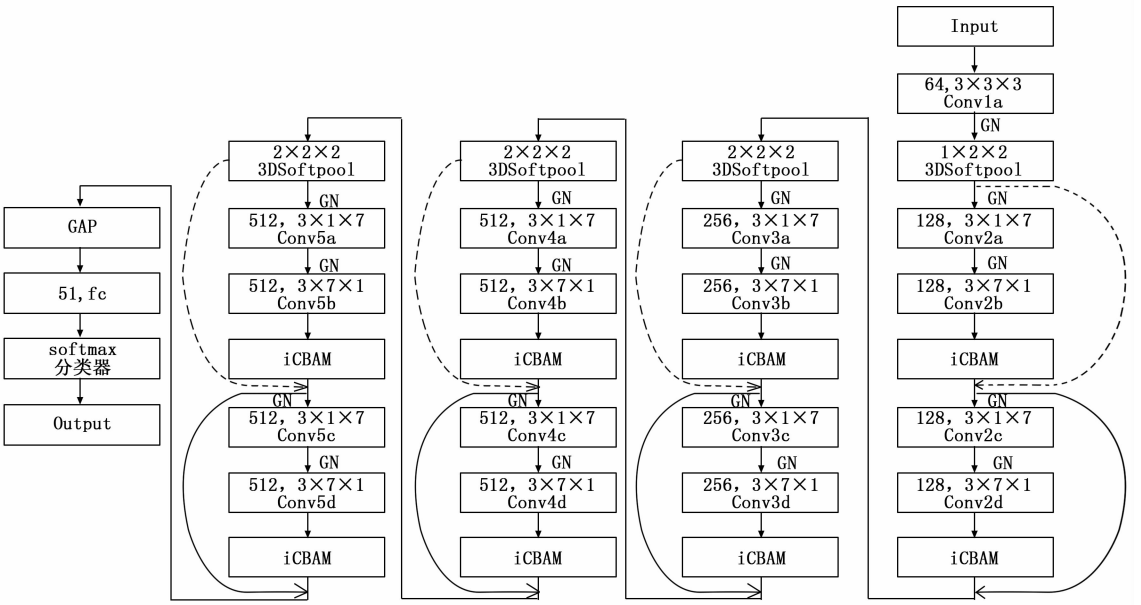

图 9 C3D 注意力残差网络结构图

的各种运动，所以无法展现本文网络结构在具体领域的应用性，为了体现本文提出的算法具有一定的应用性，则对能够产生复杂人体行为的体育运动进行识别，但是没有现成的类别数目较大的体育运动数据集，所以本文从 UCF101 数据集[31]和 kinetic400 数据集[32]中选出在比赛中常见的体育运动进行混合，则自建了具有 43 个类别的体育运动数据集。该数据集一共有 5 302 个视频片段，每个类别至少包含 108 个剪辑片段，视频中的分辨率最小为 140×256，帧率最小为 25 fps，数据集内存大小为 3.61 GB。该数据集主要涉及体操、球类、游泳、跳水、田径、滑冰、滑雪、举重等 8 类体育运动。

为评估基于 C3D 注意力残差网络模型的性能，本文采用模型行为识别准确率 ACC（Accuracy）、模型 ROC（Receiver Operating Characteristic）曲线下的面积 AUC（Area Under Curve）以及对模型 PR（Precision Recall）曲线下的面积 AP（Average Precision）值取类别总数平均得到的平均精度（mAP，mean Average Precision）在 HMDB51 数据集和体育运动数据集上对模型进行性能评价，并且为衡量模型轻量化，则采用模型参数量（Params）和浮点运算次数（FLOPs）来分别对模型的空间和时间复杂度进行评价。其中 AUC 的值能够量化地反映基于 ROC 曲线衡量出的模型性能，通常 AUC 取值区间为 [0.5，1]，并且取值越大模型的分类效果越好，同样每个类别 PR 曲线下的面积 AP，则反映了当前模型对该类别的分类性能，而 mAP 值则是对总体类别的 AP 取类别总数的平均来量化当前数据集下模型总体分类的性能，mAP 取值为 [0，1]，取值越大模型分类效果越好。由于 ROC 曲线不易受到数据样本分布的影响，而当数据样本分布相差很大时 PR 曲线更能反映分类器性能，所以为能更好的衡量模型性能，本文同时采用这两种指标对模型评价。

4.3 数据预处理

在对视频中的人体行为进行识别时需要先对视频数据进行预处理操作。首先将视频数据集按照 6：2：2 的比例形式分为训练集、验证集和测试集。之后按照原始 C3D 网络中的视频预处理方式将其以每隔 4 帧截取一帧的方式，使每个视频片段变成至少为 16 帧的连续视频帧图像，若视频帧数较少而无法满足间隔 4 帧的截取方式，则自动降低采样间隔以满足要求。得到视频帧后将视频帧图像尺寸统一调整为 171×128，再对调整后的视频帧采用随机裁剪为 112×112、以概率为 0.5 的水平翻转和去均值等数据增强操作，则最终网络的输入尺寸为（3×16×112×112）。

5 实验结果与分析

5.1 HMDB51 实验结果分析

本实验由于受到计算机显存限制，无法使用大型数据集进行预训练，所以为进行公平比较，实验中所有网络模型均在相同实验设备中从头开始训练，均没有使用任何经过大数据集训练后的预训练模型。

本文算法通过在 HMDB51 数据集上进行训练，总训练时长约为 20 h，本文算法与目前流行人体行为识别算法 C3D[8]和 Res3D[14]在 HMDB51 上进行性能比较，3 种模型的训练和测试过程如图 10 所示。

3 种模型在 HMDB51 数据集上的测试结果以及各种性能指标结果如表 1 所示。

表 1 HMDB51 数据集测试结果对比

模型	ACC/%	AUC	mAP	Params/10^6	FLOPs/10^9
C3D	31.16	0.88	0.24	78.20	38.66
Res3D-18	36.12	0.89	0.32	33.20	37.54
Res-iCBAM-C3D(本文)	41.04	0.90	0.36	47.95	45.32

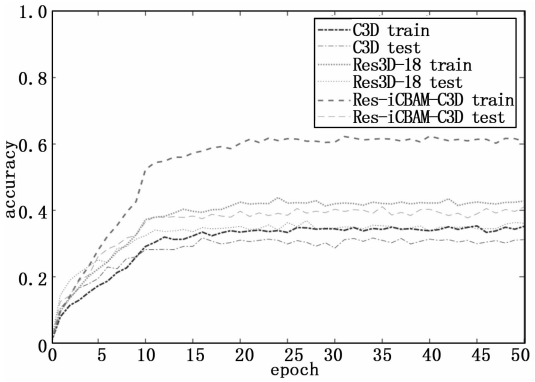

图 10　HMDB51 准确率曲线

由表 1 中的实验结果可知在没有经过任何预训练的情况下，本文算法在 HMDB51 上的识别准确率为 41.04%，该结果比 Res3D 高了 4.92%，而且比 C3D 的结果高了 9.88%，所以可知本文算法的识别效果较好，同时说明了该算法若要和 Res3D 以及 C3D 在相同大型数据集下进行预训练，则本文算法在 HMDB51 上的实验结果仍然会比这两种算法的实验结果好很多，并且会进一步提高该数据集上的识别率，所以本文从头训练模型的实验结果是有意义的。为进一步比较模型在该数据集下的识别性能，则考虑表 1 中模型的 AUC 值和 mAP 值，由于 3 个模型的 AUC 值相差不大，无法准确比较各模型性能，所以可由 mAP 值进行判断，由表 1 中模型的 AP 值大小可知，本文算法的 mAP 值为 0.36，比其余两个模型 mAP 值大，所以可知本文算法的模型性能好于其余两种模型算法。由表 1 中各模型参数和 FLOPs 可知，本文算法的参数量为 47.95 M，则本文算法比 C3D 模型的参数量降低了 38.68%，然而 FLOPs 较其余模型却有所增加，那是因为模型逐渐复杂化而造成的消极影响，但是综合考虑本文算法实现了模型轻量化，并进一步改善了模型识别效果。

5.2　体育运动数据集实验结果分析

本文为进一步展现所提出的网络结构具有一定的应用性，所以将其应用到具有 43 个类别的体育运动数据集来展示本文算法在体育运动识别方面的性能，并在该数据集中与 C3D 和 Res3D 进行比较，本文算法总训练时长约为 18.5 h，3 种模型的训练和测试过程如图 11 所示。

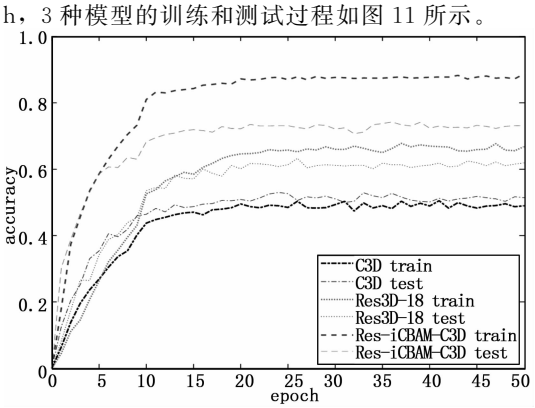

图 11　体育运动准确率曲线

3 种模型在体育运动数据集上的测试结果以及各种性能指标结果如表 2 所示。

表 2　体育运动数据集测试结果对比

模型	ACC/%	AUC	mAP	Params/10^6	FLOPs/10^9
C3D	51.44	0.94	0.46	78.20	38.66
Res3D-18	61.91	0.96	0.58	33.20	37.54
Res-iCBAM-C3D(本文)	73.05	0.97	0.68	47.95	45.32

由表 2 的实验结果可知在没有经过任何预训练的情况下，将本文算法应用到体育运动数据集中的识别结果为 73.05%，该结果比 Res3D 高了 11.14%，而且比 C3D 的结果高了 21.61%，所以可知本文算法在体育运动识别方面具有较好的识别性能。根据表 2 中模型性能的评价指标结果所示，本文算法的 AUC 值和 mAP 值分别高达 0.97 和 0.68，所以进一步说明本文网络结构性能较好，并且在体育运动识别方面具有一定的应用性。

6　结束语

本文针对 C3D 网络参数量较大以及缺少关注关键帧信息而导致识别效果不理想的问题，提出一种具有应用性的基于改进型 C3D 的注意力残差网络模型用于人体行为识别。在模型中引入非对称式三维卷积层和全局平均池化对模型进行轻量化，采用全预激活式残差结构和 iCBAM 注意力模块来提高模型的识别能力，并使用 GN 正则化和 SoftPool 进一步改善网络识别性能并加速网络训练。本文网络结构与目前流行算法在 HMDB51 数据集上进行结果对比，验证了本文方法的有效性，同时采用自建的 43 类别体育运动数据集对本文方法在实际中的应用性进行验证，结果表明本文方法同样具有良好的应用性。在未来的工作中，可以引入一些传统的特征提取方法与深度学习相结合来更好的对行为细粒度特征进行提取，以进一步增强模型的人体行为识别性能。

参考文献：

[1] LAPTEV I, LINDEBERG T. Space-time interest points [C] // Proceedings Ninth IEEE International Conference on Computer Vision. Los Alamitos, California: IEEE, 2003: 432-439.

[2] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies [C] // Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2008.

[3] RICHARDSON M, DOMINGOS P. Markov logic networks [J]. Machine Learning, 2006, 62 (1-2): 107-136.

[4] WANG H, KLASER A, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103 (1): 60-79.

[5] WANG H, SCHMID C. Action recognition with improved trajectories [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2013: 3551 - 3558.

[6] JI S, YANG M, YU K. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 221 - 231.

[7] SIMONYAN K, ZISSERMAN A. Two- stream convolutional networks for action recognition in videos [C] // Advances in Neural Information Processing Systems 27, La Jolla, California: MIT Press, 2014: 1 - 11.

[8] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C] // 2015 IEEE International Conference on Computer Vision. New York: IEEE, 2015: 4489 - 4497.

[9] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long- term recurrent convolutional networks for visual recognition and description [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 2625 - 2634.

[10] LI H, CHEN J, HU R, et al. Action recognition using visual attention with reinforcement learning [C] // International Conference on Multimedia Modeling. Berlin, Heidelberg: Springer, 2019: 365 - 376.

[11] LIU H, ZHANG L, GUAN L, et al. GFNet: a lightweight group frame network for efficient human action recognition [C] // 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2020: 2583 - 2587.

[12] YANG C Y, XU Y H, SHI J P, et al. Temporal pyramid network for action recognition [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 588 - 597.

[13] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770 - 778.

[14] TRAN D, RAY J, SHOU Z, et al. Convnet architecture search for spatiotemporal feature learning [EB/OL]. (2017- 8- 16) [2022- 1- 13]. https: //arxiv. org/abs/1708. 05038v1.

[15] QIU Z, YAO T, MEI T. Learning spatio- temporal representation with pseudo- 3d residual networks [C] // 2017 IEEE International Conference on Computer Visio. New York: IEEE, 2017: 5534 - 5542.

[16] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6450 - 6459.

[17] SHARMA S, KIROS R, SALAKHUTDINOV R. Action recognition using visual attention [C] // 2016 International Conference on Learning Representations. New York: ACM, 2016: 457 - 467.

[18] LIU Z K, TIAN Y, WANG Z L. Improving human action recognitionby temporal attention [C] // 2017 IEEE International Conference on Image Processing. New York: IEEE, 2017: 870 - 874.

[19] DAI C, LIU X G, LAI J F. Human action recognition using two- stream attention based lstm networks [J]. Applied Soft Computing, 2020, 86 (2): 202 - 209.

[20] ZHAO Z Y, ZOU W, WANG J J. Action recognition based on c3d network and adaptive keyframe extraction [C] // 2020 IEEE 6th International Conference on Computer and Communications. New York: IEEE, 2020: 2441 - 2447.

[21] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 2818 - 2826.

[22] LIN M, CHEN Q, YAN S. Network in network [C] // 2014 International Conference on Learning Representations. New York: ACM, 2014: 173 - 182.

[23] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks [C] // 14th European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2016: 630 -645.

[24] STERGIOU A, POPPE R, KALLIATAKIS G. Refining activation downsampling with softpool [EB/OL]. (2021- 3- 18) [2022- 1- 15]. https: //arxiv. org/abs/2101. 00440.

[25] WU Y X, He K M. Group normalization [C] // 15th European Conference on Computer Vision, Munich. Berlin, Heidelberg: Springer, 2018: 742 - 751.

[26] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C] // Proceedings of The 32nd International Conference on Machine Learning. New York: ACM, 2015: 448 - 456.

[27] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co- adaptation of feature detectors [J]. Computer Science, 2012, 3 (4): 212 - 223.

[28] WOO S H, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C] // 15th European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2018: 3 - 19.

[29] LIU D, JI Y F, YE M, et al. An improved attention- based spatiotemporal- stream model for action recognition in videos [J]. IEEE Access, 2020, 8: 61462 - 61470.

[30] HENDRYCkKS D, GIMPEL K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units [EB/OL]. (2016- 7- 8) [2022- 1- 16]. https://arxiv. org/abs/1606. 08415v2.

[31] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012- 12- 3) [2022- 1- 16]. https://arxiv. org/abs/1212. 0402.

[32] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset [EB/OL]. (2017- 5- 19) [2022- 1- 16]. https: //arxiv. org/abs/1705. 06950.