文章编号：2096-1472(2021)-11-51-04
DOI:10.19644/j.cnki.issn2096-1472.2021.011.013

基于注意力机制的改进残差网络的人体行为识别方法

王昊飞，李俊峰

(浙江理工大学机械与自动控制学院，浙江 杭州 310018)
✉haofeiwang@yeah.net; ljf2003zz@163.com

摘　要：针对ResNeXt网络(残差网络)中存在的对特征提取不充分，以及数据集中背景信息干扰的问题，将ResNeXt网络和注意力机制相结合，提出了一种基于注意力机制的ResNeXt模型。首先，在ResNeXt网络的基础上，将浅层和深层的特征融合生成新型网络结构。其次，将全连接层由全局平均池化层替代，然后在通道空间注意力机制中添加一个条件因子，同时将改进后的注意力机制嵌入上述网络中。最后，在UCF101和HMDB51上分别进行实验，得到了95.2%和65.6%的准确率。研究表明，本文提出的模型可以有效地提取关键特征，充分利用不同层次的特征信息获得较好的准确率。

关键词：人体行为识别；注意力机制；ResNeXt；全局平均池化
中图分类号：TP183　**文献标识码**：A

Human Action Recognition Method based on Attention Mechanism and Improved ResNeXt Network

WANG Haofei, LI Junfeng

(*Faculty of Mechanical Engineering & Automation*, *Zhejiang Sci-Tech University*, *Hangzhou* 310018, *China*)
✉haofeiwang@yeah.net; ljf2003zz@163.com

Abstract: Aiming at problems of insufficient feature extraction in ResNeXt network and background information interference in the dataset, this paper proposes a ResNeXt model based on attention mechanism, which combines the ResNeXt network and attention mechanism. First, based on ResNeXt network, shallow and deep features are merged to generate a new network structure. Second, the fully connected layer is replaced by a global average pooling layer. Then channel attention mechanism is improved by adding a condition factor. At the same time, the improved attention mechanism is embedded in the above-mentioned network. Finally, experiments are performed on UCF101 and HMDB51 respectively, and the accuracy rates of 95.2% and 65.6% are obtained. Experiments show that the proposed model can effectively extract key features, and make full use of feature information of different layers to achieve better accuracy.

Keywords: human action recognition; attention mechanism; ResNeXt network; global average pooling

1 引 言(Introduction)

人体行为识别技术是从包含运动信息的图像、视频中进行识别的。在视频监控、智能家居、运动分析以及VR等领域都离不开人体行为的识别。人体行为识别已成为计算机视觉研究中的一个非常重要的领域[1]。由于视点的不同、背景的复杂性以及光照条件等的影响，人体行为识别仍然是一项非常具有挑战性的课题。传统人体行为识别是基于手工设计的特征[2]进行识别，并且依赖数据集特征提取的先验知识，耗费大量的时间和精力。随着深度学习的兴起，解决了手动设计特征的不足，在人体行为识别领域取得了重大进展[3]，已经明显

超过了手工设计的特征。XIE等[4]提出了ResNeXt网络，用一种平行堆叠相同拓扑结构的blocks来代替残差网络三层卷积的block，同时增加了“基数”这一概念，减少了超参数数量，计算效率高，准确率高。注意力机制可以将其他不重要的信息忽略掉，重点关注关键信息[5]。将注意力机制应用到视频中的行为识别，能够有效提取视频帧中的关键信息。基于上述方法，为了充分提取视频中的特征，本文对ResNeXt网络进行改进并嵌入了改进后的通道空间注意力机制模型。

2 改进后的ResNeXt网络结构(Improved ResNeXt network architecture)

首先，本文将使用改进后的ResNeXt网络作为特征提取网络来提取时空特征，并将不同层次的特征进行融合，以充分利用各类特征信息。其次，网络中嵌入改进后的通道空间注意力机制，使网络更加关注有强反馈能力的特征。最后，经过全局平均池化操作后送入softmax函数进行分类，得到最终结果。本文提出的网络结构如图1所示。

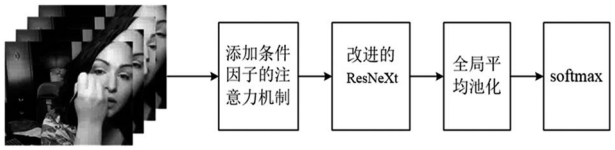


图1 网络结构

Fig.1 Network architecture

本文采用的卷积网络为ResNeXt101，主体由四个残差模块组成。残差模块的结构如图2所示，1×1×1和3×3×3表示卷积核大小，*F*表示通道数，*group*表示分组卷积的组数，即将特征图分成*group*组的小特征图。ResNeXt网络结构采用VGG网络和inception网络中转换合并的思想，用一种平行的相同拓扑结构的block进行堆叠来进行分组卷积，用来控制分组数量，在没有增加参数复杂度的情况下提高了准确率。

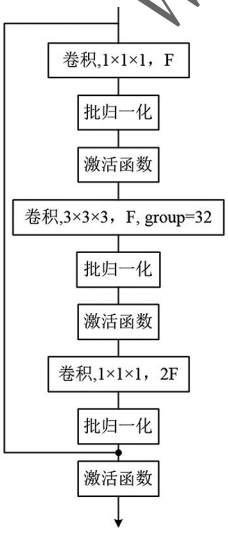


图2 残差模块

Fig.2 Residual module

本文中网络的残差模块分别用layer1、layer2、layer3、layer4表示，网络的具体结构如图3所示。随着网络的加深，一些细节特征被过滤掉，导致对提取到的特征利用不充分。本文改进后的ResNeXt网络将浅层网络提取的细节特征和深层网络提取的特征相融合，以充分利用各个层次所提取的特征信息。

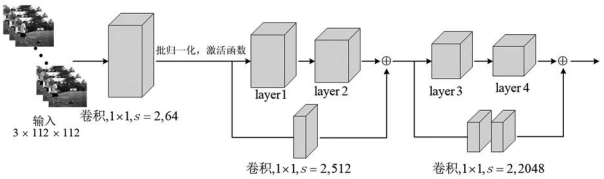


图3 改进后的ResNeXt网络结构

Fig.3 Improved ResNeXt network architecture

对注意力机制输出的特征进行步长为2、卷积核大小为1的卷积，卷积操作后的特征和layer2输出的特征相融合输入layer3中继续进行卷积操作。同理，将和layer2融合后的特征进行两次步长为2、卷积核大小为1的卷积操作，并和layer4输出的特征相融合。进行卷积操作的目的是为了降低维度，使特征图能够进行融合。文中没有采用逐层特征融合，而是采用跳层融合的方式，首先是为了降低模型参数，减少计算量；其次，如果采用逐层融合的方式，包含过多的特征，会造成冗余的信息。两种特征采用element−wise进行融合。

3 注意力机制(Attention mechanism)

注意力机制模型[6]如图4所示，由通道注意力机制和空间注意力机制串联组成，对特征图在通道和空间维度上进行注意力生成，可以在不明显增加计算量的基础上提高准确率。

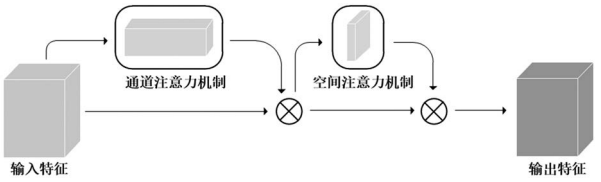


图4 注意力机制模型

Fig.4 Model of attention mechanism

(1)通道注意力机制

本文对通道注意力机制进行了改进，由于平均池化和最大池化提取到的特征有所区别，添加了条件因子∂来对不同的特征进行权重分配。改进后的通道注意力机制如图5所示。首先将输入特征图在空间维度上进行压缩，分别进行平均池化和最大池化操作，得到F^c_{avg}和F^c_{max}。然后对得到的这两个特征图进行权重分配，将这两个重新分配的特征输入一个共享网络中，该共享网络是包含一个隐藏层的多层感知机(MLP)，经过共享网络的处理后，用element−wise求和输出特征向量。

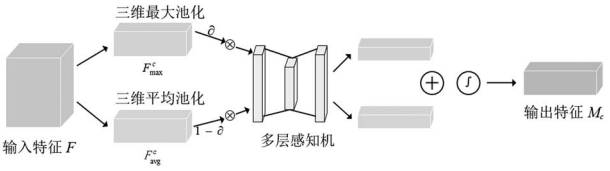


图5 通道注意力机制

Fig.5 Channel attention mechanism

计算公式如式(1)所示：

$$\boldsymbol{M}_c = \sigma\left(\mathrm{MLP}\left(\partial\left(\mathrm{Avgpool}\left(\boldsymbol{F}\right)\right)\right) + \mathrm{MLP}\left(\left(1-\partial\right)\left(\mathrm{Maxpool}\left(\boldsymbol{F}\right)\right)\right)\right) \quad (1)$$

其中，F为输入特征向量，MLP为多层感知机的处理过程，Avgpool为平均池化，Maxpool为最大池化，∂为条件因子，σ为sigmoid函数，$\boldsymbol{M}_c$为输出特征向量。

(2)空间注意力机制

空间注意力机制如图6所示，将特征图在通道维度上进行压缩。对输入的特征图分别在通道维度做平均池化和最大池化操作，得到两个二维特征；然后，按照通道将特征进行拼接得到一个特征图；最后，对其进行卷积操作，使得最终得到的特征图和输入的特征图在空间维度上一致。

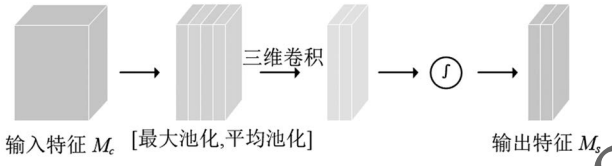


图6 空间注意力机制

Fig.6 Spatial attention mechanism

计算公式如式(2)所示：

$$\boldsymbol{M}_s = \sigma\left(f^{7\times7}\left(\left[\mathrm{Avgpool}\left(\boldsymbol{F}\right);\mathrm{Maxpool}\left(\boldsymbol{F}\right)\right]\right)\right) \quad (2)$$

其中，$\boldsymbol{M}_s$为输出特征向量；σ为sigmoid函数；f为卷积计算；Avgpool为平均池化；Maxpool为最大池化。

4 全局平均池化(Global average pooling)

传统的卷积神经网络分类时使用全连接层和softmax回归层。但是，由于全连接层参数过多，计算量大，容易造成过拟合，同时全连接层容易导致特征图损失空间位置信息。因此，本文采用全局平均池化层[7]来代替ResNeXt的全连接层，使特征图和行为类别之间的联系更加直观，转换为分类的概率更加容易，对空间位置信息的鲁棒性更强。

全局平均池化是对每一个通道图的所有像素求平均值，在特征提取的最后一个卷积层生成k个特征图；经过全局平均池化层后得到k个1×1的特征图，将这些特征图输入softmax层，输出结果就是k个类别的置信度。

图7为全局平均池化示意图，图8为全连接示意图。本文对图7和图8进行参数计算，假设输入特征图大小为$3\times3\times3$，则全连接层产生的参数个数为$3\times3\times3\times3=81$个，而全局平均池化层将输入特征进行池化后直接送入softmax，所以参数个数为$3\times1\times1\times3=9$个。相比于全连接层，全局平均池化层的参数成倍数减少。

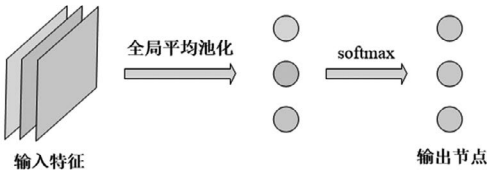


图7 全局平均池化

Fig.7 Global average pooling

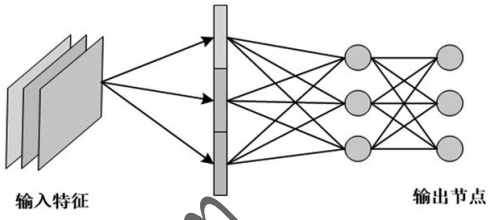


图8 全连接

Fig.8 Fully connected

5 实验(Experiment)

5.1 数据集

(1)UCF101数据集

UCF101[8]是行为类别和样本数量最多的数据库之一，其中包含13,320 个视频和101 个类别。数据库的样本取自从BBC/ESPN收集并从网络上下载的各种运动的样本。UCF101多样性较强，在相机运动，人体的外形、形态、视点、背景、光照条件等各种不同的条件下存在较大差异，是目前为止最具挑战性的数据库之一。101 类行为被分成25 组，每组包括4—7 个视频，主要分为人与物体之间的交互、人与人之间的交互、人体自身的行为、演奏乐器和运动五类，如画眼妆、打篮球、打太极拳、弹吉他、攀岩等。同一组视频可能有一些共同的特征，如背景、视点等。如图9所示为部分动作示意图。

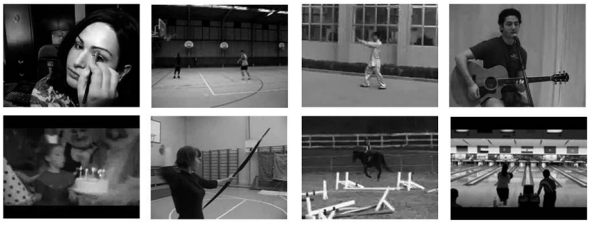


图9 UCF101数据集部分动作示意图

Fig.9 Schematic diagram of some of the actions in the UCF101 dataset

(2)HMDB51数据集

HMDB51[9]包含6,849 个视频，总共51 个类别，每个类别至少包含101 个视频。大多数视频来自电影片段，有些来自公共数据库，例如YouTube。动作主要包含一般面部的行为、面部的操作与对象的操作、身体的行为、身体与对象交互的

行为和人体自身的行为五类，如交谈、喝水、倒立、骑自行车、拥抱等。部分动作示意图如图10所示。

图10 HMDB51数据集部分动作示意图

Fig.10 Schematic diagram of some of the actions in the HMDB51 dataset

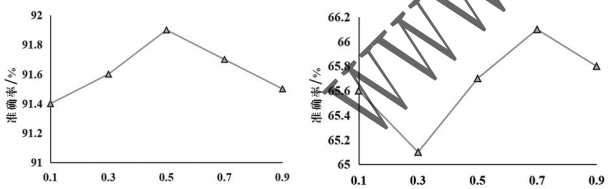
5.2 视频采样与参数设置

本文将视频随机的一个位置进行均匀采样生成16 帧的输入片段，并通过裁剪的方式将样本尺寸统一为112×112，所以网络的输入样本大小为3×16×112×112。训练过程中，初始学习率设置为0.05，并在验证损失达到饱和后将其除以10，进行学习率衰减优化。使用动量为0.9的随机梯度下降优化器来对网络进行优化，使用ReLU激活函数，采用交叉熵损失函数计算损失。

5.3 结果与分析

(1)不同条件因子下的比较实验

该部分就改进的注意力机制中的条件因子的不同取值进行实验，分别在UCF101和HMDB51数据集划分的spilt1部分进行实验，条件因子∂分别取0.1、0.3、0.5、0.7、0.9，得到的结果如图11所示。可以看出，在UCF101上，当∂取值为0.5时，效果较好；在HMDB51上，当∂取值为0.7时，效果较好。所以本文选取∂为0.5和0.7分别进行实验。



(a)UCF101数据集 (b)HMDB51数据集

图11 不同∂下的准确率

Fig.11 Accuracy with different ∂

(2)拆分实验

该部分将数据集UCF101和HMDB51分别拆分成三个部分进行实验，取三者的平均值作为最终结果。UCF101数据集被分成三个部分，每个部分包含测试集和训练集，每类行为的测试集和训练集总共为25 组，其中测试集包含7 组，训练集包含18 组。三种不同的拆分方式中的测试集交叉取前中后7 组，训练集取剩下的18 组，三个部分的测试集和训练集一一对应。而HMDB51数据集随机生成三种拆分方式。首先选择元标签分布最平衡的片段，然后选择与之关联最小的第二、第三片段，一次得到三种不同的拆分方式。每种拆分方式的每类行为都包含70 组训练片段和30 组测试片段，结果如表1所示。

表1 三个拆分方式的实验结果

Tab.1 Experimental results of three splits methods

划分类别	UCF101数据集	HMDB51数据集
Spilt1	91.9%	66.1%
Spilt2	96.8%	64.4%
Spilt3	96.8%	66.3%
Average	95.2%	65.6%

(3)有无注意力机制对比实验

该部分对添加了注意力机制的特征图进行了可视化，将生成的热力图和原图相结合，如图12所示。图中热力图深色区域表示所预测到的行为，浅色区域表示背景部分，深色越深代表所受的关注越多。可以看出，添加注意力机制模型后，能够更有效地集中在关键信息处，能够更好地提取行为的关键信息，以便提高识别的准确率。本部分有无注意力机制模型进行对比的实验结果如表2所示。由表2可知，添加注意力机制后，无论是在UCF101还是在HMDB51上的准确率都有一定的提升。

图12 注意力机制可视化

Fig.12 Visualization of attention mechanism

表2 有无注意力机制的实验结果

Tab.2 Experimental results with and without attention mechanism

是否有注意力机制	UCF101数据集	HMDB51数据集
无	92.8%	65.2%
有	95.2%	65.6%

(4)与其他算法的对比实验

为了验证本文算法的有效性，在数据集UCF101和HMDB51上，与近年来主流的iDT[10]、TSN[11]、Two-Stream CNN[12]等人体行为识别方法进行了比较，实验结果如表3所示。结果表明，本文的识别模型相比一些主流模型准确率有了大幅度提高，尤其在UCF101数据集上比iDT、Two-Stream分别提高了8.8%和7.2%。

(下转第46页)