文章编号：2095-2163(2021)12-0153-05 中图分类号：TP391.4 文献标志码：A

基于神经网络和迁移学习的视频人体行为识别

吴松平，王天一

（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：为了解决视频人体行为识别中网络难以训练、直接将卷积神经网络全连接层的输出送入循环神经网络而导致空间信息缺失，进而引起视频人体行为识别精度不高、难以训练等问题。本文提出基于神经网络和迁移学习的视频人体行为识别方法，该方法以 resnet50 为基础网络，将在 imagenet 数据集上训练好的权重参数用于初始化所有的卷积层，使用卷积长短期记忆神经网络对 resnet50 的输出做处理，得到具有空间信息的视频描述信息，使用注意力机制对视频信息进行处理得到视频关键信息，最后利用长短期记忆网络对视频关键信息做时间序列建模。该方法在人体行为通用数据集 ucf101 上到达 94.77%。经实验证明，该方法可以实现端到端的视频人体行为识别，识别精度可以和现有的方法媲美，并有训练时长短，网络结构简单等特点。

关键词：人体行为识别；卷积长短期记忆神经网络；注意力机制；ucf101

Video human behavior recognition based on neural network and transfer learning

WU Songping, WANG Tianyi

(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 In order to solve the problem that the network is difficult to train in the video human behavior recognition and the lack of information caused by directly sending the output of the fully connected layer of the convolutional neural network to the recurrent neural network, this paper proposes a video human behavior recognition method based on neural network and transfer learning. This method uses resnet50 as the basic network whose weight parameters are trained on the imagenet dataset to initialize all convolutional layers, and uses the convolutional long and short term memory network to process the output of resnet50 to obtain a video description with spatial information. The attention mechanism is also used to process the video information to obtain the key video information, and finally the long and short term memory network is used to model the time series of the key video information. This method achieves a recognition accuracy of 94.77% on the universal human behavior data set ucf101. Experiments have proved that this method can realize end-to-end video human behavior recognition. The recognition accuracy can be comparable to existing methods, and it has the characteristics of short training time and simple network structure.

【Key words】 human behavior recognition; convolutional long and short term memory neural network; attention mechanism; ucf101

0 引 言

大数据、人工智能的快速发展，产生了大量的视频数据，对这些视频数据进行智能分析、视频摘要、视频信息检索、运动分析等方面有重要的意义[1]。行为识别作为视频分析的一个重要领域，相比传统人体行为识别方式[2]。基于深度学习的人体行为识别方法能够实现端到端的识别，深度学习算法的研究推动了行为识别研究的进步。

基于深度学习的人体行为识别的基本原理是通过构建具有提取非线性特征的卷积神经网络、具有时间序列建模的循环神经网络结构，并利用数据对网络进行训练，得到效果最好的网络参数，提取数据集与人体行为最相关的本质特征。目前，相对于在图像分类、人脸识别、图像分割等方面的任务中的表现，深度学习在人体行为识别的表现依然欠佳，其原因在于视频数据相比图像数据还多了时间维度信息[3]。一般的神经网络只能处理静态图像数据，而不能充分利用人体运动信息[4]。双流（Two-Stream）卷积神经网络是目前运用最广泛的方法之一，与主要依靠图像数据信息进行视频分析的传统方法相比，双流卷积神经网络在以图像数据为信息的基础上加入了时间光流信息，两种信息分别送入卷积神经网络，最后将两路信息进行特征融合[5]。在双流卷积神经网络的基础上，Du Tran 等人提出了 3D 卷积，将 2D 卷积核替换成 3D 卷积核，直接对

作者简介：吴松平（1996-），女，硕士研究生，主要研究方向：深度学习、图像识别；王天一（1989-），男，博士，副教授，主要研究方向：量子通信、图像处理、计算机视觉。

通讯作者：王天一 Email：tywang@gzu.edu.cn

收稿日期：2021-09-12

视频帧处理[6];Feichenhofer 等人探索了双流卷积神经网络的融合时机[7]。

以上方法虽然综合了时间光流信息和图像信息,取得了较高的识别精度,却付出了复杂度的代价,同时对长时间人体行为分析往往并不准确[8]。循环神经网络对长时序列建模有很好的效果,但对长时间的时间建模容易产生梯度爆炸和梯度消失等问题。长短期记忆神经网络能够解决循环神经中出现的梯度爆炸和梯度消失的问题,被广泛用于机器翻译、语言识别等具有时间序列的任务中[9]。视频数据相较于图片数据来说具有时间先后顺序的属性,Donahue 等人提出将长短期记忆神经网络应用于视频数据的描述与识别,取得了较好的识别效果[10]。一般意义上的长短期记忆循环神经网络只能处理一维向量数据,视频帧经过特征提取网络得到特征图,在进入长短期记忆循环神经网络前都要把数据降维成一维向量数据,该操作将使数据丧失空间特征。

为解决以上问题,本文将卷积长短期记忆神经网络运用到视频行为识别中,在对视频帧进行时间序列建模的同时,还能够兼顾空间信息。本文先利用卷积长短期记忆网络对基础网络提取到的视频帧特征进行一次时间序列建模,得到具有空间信息的视频描述,对视频描述进行下采样,将下采样结果送入长短期记忆神经网络进行二次时间序列建模。得到识别效果能够与现存算法相媲美。

1 识别网络

1.1 识别网络图

本文识别网络如图 1 所示。对输入的视频帧先做特征提取,得到视频帧中层特征,对中层特征做兼顾空间信息的初次时间序列建模,获得初级视频描述,将视频描述做注意力操作。

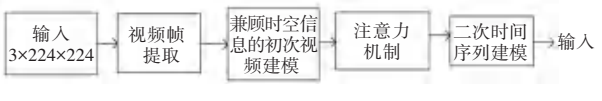

图 1 识别网络图

Fig. 1 Identification network diagram

1.2 视频帧提取模块

该部分使用 resnet50 网络提取特征,残差网络相对于其他神经网络能够解决梯度消失、梯度爆炸等问题。

resnet50 的网络参数,见表 1。resnet50 卷积核有 7×7、3×3、1×1 3 种,resnet50 网络有 5 个部分,除了第一个部分由卷积核为 7×7 卷积层和 3×3 最大池化层组成,其他的部分由多个残差单元组成[11]。随着卷积层一层层的运算,卷积核输出的内容越来越抽象,保留的空间信息也越来越少,卷积层后经过平均池化操作减少特征图的尺寸。平均池化完成后将特征图打平为一维数据,作为全连接层的输入,连接层由1 000个神经元组成。

本文使用除了最后全部连接层以外其余部分作为基础特征提取网络。基础网络对视频帧做特征提取,通过网络训练提取到利于识别任务的中高层特征。

表 1 resnet50 网络参数表

Tab. 1 resnet50 network parameters

Layer name	Output size	Layer
Conv1_x	112×112	7×7,64,stride2 3×3 max pool stride2
Conv2_x	56×56	[1×1,64; 3×3,64; 1×1,256]×3
Conv3_x	28×28	[1×1,128; 3×3,128; 1×1,512]×4
Conv4_x	14×14	[1×1,256; 3×3,256; 1×1,512]×6
Conv5_x	7×7	[1×1,512; 3×3,512; 1×1,2 048]×3
Average pool.10000-d fc .softmax		

1.3 兼顾时空信息的初次时间序列建模

为充分利用人体行为视频中的时空信息,该部分使用卷积长短期记忆神经网络。与长短期记忆神经网络相比,保留了长短时记忆神经网络的优点的同时,还可以处理视频帧的空间信息[12]。卷积长短期记忆神经网络与长短期记忆神经网络相比在运算公式上有所不同,卷积长短期记忆神经网络的运算公式是在长短期记忆神经网络上的改进,把长短期记忆神经网络中的乘法运算换成卷积运算,因此能够保留空间信息同时做时间序列建模。

长短期记忆神经网络只能够对一维数据做时间序列建模,对图像数据做时间建模时,必须将图像数据处理为一维数据,处理过程使图像数据失去空间信息。卷积长短期记忆神经网络处理数据时,不必打平具有位置信息的视频帧图像,在保留空间信息

的前提下做时间序列建模。该网络运用多维度信息比使用单维信息在识别效果上有很大的改进。

1.4　注意力机制

视觉注意力机制是人眼看到物体时的信息处理过程。在观察一个物体时，人眼和大脑会自动给重要的特征更多的注意力，对于不同的物体，注意力的中心区域也会发生变化[13]。视频图像帧中有很多背景信息，背景信息会给识别过程带来干扰，引入注意力机制有助于去除干扰信息、提高识别精度。

本文使用的注意力机制原理图如图 2 所示。平均注意力机制，对空间位置取平均，并失去空间位置。通道注意力机制对上一个模块的输出在通道维度进行平均值操作，在空间上进行卷积操作，通过最大激活函数在空间维度上计算，得到注意力权重图，最后将注意力权重图运用到特征图中。线性注意力机制是使用线性操作对通道维度进行处理，得到注意力图，将注意力图和原始特征图相乘得到最终注意力机制的特征图。将 3 种注意力机制的输出分别经过长短期记忆神经网络做二次时间序列建模，将二次时间序列建模结果送入全连接层，最终将全连接层的输出取平均值，得到最终的输出结果。

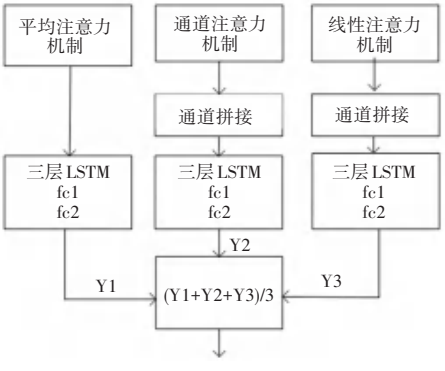

图 2　注意力机制原理图

Fig. 2　Schematic diagram of attention mechanism

1.5　迁移学习

迁移学习的目标是把源域学习到的信息应用到目标域中。在深度学习中，训练数据决定模型的训练效果[14]。然而，在大多数任务中，缺乏训练数据，导致识别的效果不理想。另外，大量没有进行标注的数据样本，无法直接使用，需要耗费大量的人力去标注。而迁移学习可以解决训练数据缺乏、数据标注难度大等问题。将在相似数据集上训练得到的网络权重迁移到目标网络，能够更快更好地进行参数的训练，而不必从头训练。

本文采用的迁移方式使用 imagenet 数据集的权重参数，冻结全连接层之前的所有权重参数，进行其他参数的训练。

2　基于迁移学习的神经网络模型

2.1　数据集以及图像预处理

本文采用公开数据集 UCF101，包含13 320个视频(共 27 h)，利用 OpenCV 对 UCF101 中的视频保持结构不变，逐帧分解得到图像，UCF101 主要包括 5 大类动作，人与物体交互，单纯的肢体动作，人与人交互，演奏乐器，体育运动。该视频数据集是行为识别领域较常用的通用数据集，由于视频帧中对最终分类任务有效果的只是极少数的视频帧图像，大多数视频帧对最终识别任务是无效的，因此对视频帧采样，可以在减少训练时间的同时达到较为理想的识别效果。利用 OpenCV 来做视频预处理得到三通道图像视频帧，将视频帧的大小裁剪为 224×224，送入基本特征提取网络中高层特征。

考虑到人体行为识别并不一定要用到视频中所有的帧，本论文截取每个视频的前 40 帧代表每个视频；相邻视频帧之间的特征差别并不大，为了找到效果最好的视频间隔，本论文在 40 帧之间分别隔 2 帧、隔 4 帧、隔 6 帧、隔 7 帧采样进行实验。

2.2　神经网络深度迁移模型

本文使用 imagenet 数据集上训练得到的参数来初始化基本网络，并冻结该网络。虽然这些参数并不是由人体行为数据集图像训练得到的，但是人体行为识别图像都是普通的图像，没有特别难以理解图像，这些参数对该视频数据图像能够兼容，理论上此迁移学习会有很好的结果。

将 resnet50 网络最后的平均池化层、全连接层和激活函数去掉，得到基本网络。在基本网络后连接卷积长短期记忆神经网络，对基本网络的输出做具有空间信息的时间序列建模，得到具有空间、时间信息的视频描述。用注意力机制对卷积长短期记忆神经网络的输出做注意力操作，对输出特征图打平操作得到一维数据，一维数据送入长短期记忆神经网络，对一维数据进行二次时间序列建模。

迁移识别模型如图 3 所示。基本网络使用在 imagenet 数据集上，训练参数来训练人体行为识别视频数据集的三维图像帧。卷积长短期记忆神经网络、注意力机制、长短期记忆神经网络、全连接层使用随机参数初始化方法。卷积长短期记忆神经网络使用一层网络结构做第一次带空间信息的时间序列建模，使用 512 个卷积核对特征图卷积，特征图空间大小不变。长短期记忆神经网络使用三层网络结构对打平

后的数据做二次时间序列建模,并同时将数据维度减半,取最后一个时间步骤的输出作为视频高级特征,经过二层全连接层得到最后的视频描述。

视频图像帧为 3 特征通道,像素大小 224×224。图像特征图大小刚好与预训练参数模型大小相符合。

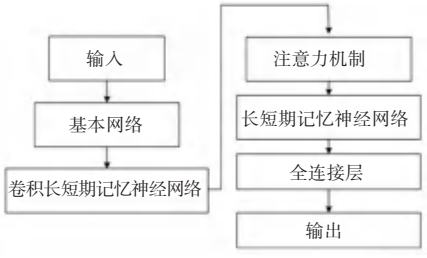

图 3 迁移识别模型

Fig. 3 Migration identification model

3 实验结果分析

本论文实验均在 Ubuntu18.04.4LTS 操作系统上进行,采用 Pytorch 深度学习框架,i7-9700 处理器,RTX2080Ti 显卡,32GB 内存。实验采用 resnet50 模型作为基本特征提取网络,损失函数默认使用 Cross Entropy Loss,Adam 优化器,Batch Size 为 30,学习率为 1×10^{-5},每组实验训练 120 个 epochs。本论文将 UCF101 视频数据集按 3:1 的比例划分为训练集和测试集,即将 UCF101 的 13 320 个视频中的 9 990个视频数据作为训练集,3 330 个视频数据作为测试集。

3.1 在神经网络上的识别效果

为验证本文方法,做 3 组实验:

(1)在基本网络后连接卷积长短期记忆神经网络;

(2)在基本网络后连接长短期记忆神经网络;

(3)在基本网络后连接卷积长短期记忆神经网络、注意力机制、长短期记忆神经网络。

使用基本网络的基础上加上卷积长短期记忆神经网络、注意力机制、长短期记忆神经网络的训练结果对比图如图 4 所示,横坐标为训练次数,纵坐标分别为准确率和损失值。可以看出,无论是准确率还是损失值,相比于其他网络,基础网络加上卷积长短期记忆神经网络和长短期记忆神经网络收敛的更快且更加平稳。

图 4(a)中绿色曲线为使用卷积长短期记忆神经网络和长短期记忆神经网络双重时间序列建模的测试集识别精度曲线,曲线显示在第 36 个 epochs 以后趋于平稳,此时的识别精度为 93.39%,在第 108 个 epochs 时达到最高识别精度 94.77%;橙色曲线为使用长短期记忆神经网络而没有使用卷积长短期记忆神经网络的测试集识别精度曲线,该识别精度曲线在第 92 个 epochs 以后趋于平稳,此时的识别精度为 88.01%,在第 120 个 epochs 时达到最好识别精度;蓝色曲线为使用卷积长短期记忆神经网络而没有使用长短期记忆神经网络的识别精度曲线,可以看到该曲线明显比其他曲线差,没有明显的收敛;通过对比双重时间序列建模模型使得人体行为识别较快的到达收敛,识别精度也比单重时间序列建模高出几个点,绿色曲线虽然在几个时间点波动较大,但是其始终在其他曲线的上方,总体效果比其他曲线理想。

在图 4(b)中,绿色曲线为使用卷积长短期记忆神经网路和长短期记忆神经网路双重时间序列建模的损失曲线,可以看出,绿色曲线在收敛速度、波动幅度上都比只使用长短期记忆神经网络和只使用卷积长短期记忆神经网络的损失曲线效果好。绿色曲线在第 40 个 epochs 以后趋于平稳,而橙色曲线在第 100 个 epochs 以后才趋于平稳。和识别精度曲线一样,双重曲线在几个时间点上波动较大,但绝大多数都在单重曲线下方。

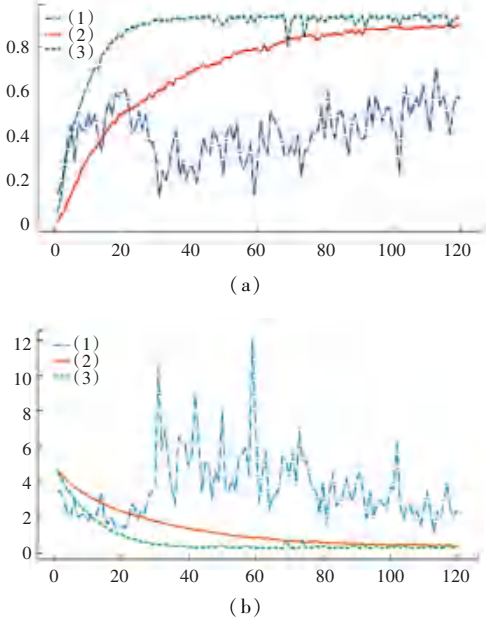

图 4 神经网络识别结果

Fig. 4 Recognition results

3.2 基于迁移的神经网络不同采样间隔识别结果

视频是由一张张图像帧组成,相邻的视频帧相似度很高,视频帧全部送入网络将带来时间复杂度。对视频图像帧采样能够兼顾时间和精确度,对视频

帧不同的采样帧的实验结果如图 5 所示。

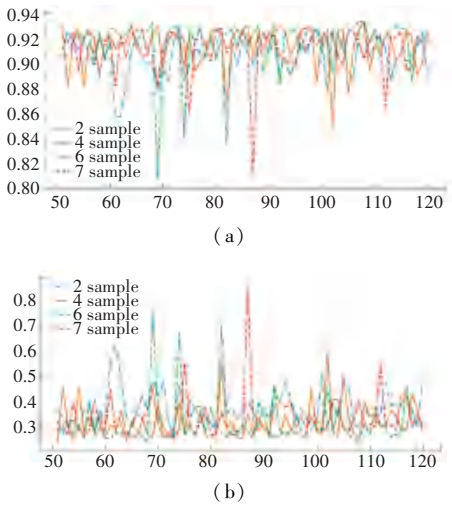

图 5　不同采样帧数的识别结果

Fig. 5　Recognition results of different sampling frames

图 5(a)和图 5(b)分别是视频帧不同的采样帧在精度和损失两方面的实验结果。隔 6 帧采样的识别精度曲线大部分区域都是在其他曲线的上方，而损失曲线的大部分都是在其他曲线的下方。虽然识别效果不是很明显，在识别精度略有提高的同时训练时间上有很大优势。

间隔 6 帧采样能够兼顾精度和时间的要求，采样帧数越大，每个视频采样到的视频帧数越小，训练所需要的时间越少，见表 2。在间隔 6 帧之前识别精度都在前列的基础上略有提高，训练所需要花费的时间逐渐减小，而在间隔 7 帧采样时，识别精度开始有明显下降，其主要原因是间隔太密集视频帧有很多相似的空间特征，使得在人体行为识别过程中错误识别为其他行为，识别率较低；间隔帧数超过 6 帧以后，视频帧空间特征相似度减小，丧失有区别的空间特征，网络不能够提取关键信息，以至于识别效果下降。

表 2　不同采样间隔的精度和时间表

Tab. 2　Accuracy and schedule of different sampling intervals

采样间隔	精度/%	时间/h
隔 2	94.20	37
隔 4	94.59	18
隔 6	94.77	13.5
隔 7	94.05	12

4　结束语

目前，深度学习模型都依赖大量的训练数据，数据量不够会出现网络无法训练或者欠拟合等问题。本文采用 resnet50 的前 49 层网络作为基础网络，结合迁移学习的方法提取视频帧的基本特征，将得到视频帧特征送入卷积长短期记忆神经网络进行兼顾空间信息的第一时间序列建模；将得到的视频描述在空间上进行下采样，得到丧失空间信息的视频帧特征；最后送入普通长短期记忆神经网络做二次时间序列建模。对视频数据集进行提帧和采样工作，隔 6 帧采样能够兼容训练时间和识别精确度。

参考文献

[1] 朱煜，赵江坤，王逸宁，等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报，2016，42(6)：848-857.

[2] 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.

[3] 李庆辉，李艾华，王涛，等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报，2018，38(6)：234-240.

[4] YAO G，LEI T，ZHONG J. A review of Convolutional-Neural-Network - based action recognition [J]. Pattern Recognition Letters，2019，118：14-22.

[5] SIMONYAN K，ZISSERMAN A. Two - stream convolutional networks for action recognition in videos[J]. arXiv preprint arXiv：1406.2199，2014.

[6] TRAN D，BOURDEV L，FERGUS R，et al. Learning Spatiotemporal Features with 3D Convolutional Networks[J]. Proceeding of the IEEE International Conference on Computer Vision，2015：4489-4497.

[7] FEICHTENHOFER C，PINZ A，ZISSERMAN A. Convolutional Two-Stream Network Fusion for Video Action Recognition[J]. Preceeding of the IEEE Conference on Computer Vision and Pattern Recognition，2016：1933-1941.

[8] 揭志浩，曾明如，周鑫恒，等. 结合 Attention-ConvLSTM 的双流卷积行为识别[J]. 小型微型计算机系统，2021，42(2)：405-408.

[9] HOCHREITER S，SCHMIDHUBER J. Long short-term memory [J]. Neural Computation，1997，9(8)：1735-1780.

[10] Donahue J，Hendricks L A，Rohrbach M，et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Trans Pattern Anal Mach Intell，2017，39(4)：677-691.

[11] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[J]. Proceedings of the IEEE conference on computer vision and pattern recognition，2016：770-778.

[12] XINGJIAN S H I，CHEN Z，WANG H，et al. Convolutional LSTM net- work：a machine learning approach for precipitation nowcasting [C]//Advances in Neural Information Processing Systems，2015.

[13] 王增强，张文强，张良. 引入高阶注意力机制的人体行为识别[J]. 信号处理，2020，36(8)：1272-1279.

[14] PENG Y，WANG S，LU B L. Marginalized denoising autoencoder via graph regularization for domain adaptation[C]//International Conference on Neural Information Processing. Springer，Berlin，Heidelberg，2013：156-163.