第 35 卷 第 6 期
2020 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 35 No. 6
Dec. 2020

DOI:10.13364/j.issn.1672-6510.20190252
数字出版日期：2020–07–06；数字出版网址：http://kns.cnki.net/kcms/detail/12.1355.N.20200704.1030.002.html

基于 ResNet-LSTM 的具有注意力机制的办公人员行为视频识别

张传雷，武大硕，向启怀，陈 佳，刘丽欣
（天津科技大学人工智能学院，天津 300457）

摘 要：针对人体行为视频分析中时间域和空间域上特征提取存在的易混乱以及运算成本高的问题，提出了 ResNet-LSTM-Attention 网络模型结构. 该网络结构将空间域和时间域的特征分开提取，对于空间维度，将单帧图像输入残差网络（ResNet）模型，提取空间维度特征；对于时间维度，将多帧叠加后的空间维度特征作为输入，输入循环神经网络（LSTM）和注意力（Attention）机制的融合模型；最后将融合模型的输出经过 softmax 作为识别结果. 将本网络模型应用到实验室采集的办公人员行为数据集上进行人体行为识别实验，并与视频分析中常用的 C3D 网络模型和无注意力机制的 ResNet-LSTM 模型进行了对比. 实验结果表明，本文模型不仅有较高的识别准确率，而且训练用时和运算成本也大大降低.

关键词：人体行为识别；视频分析；残差网络（ResNet）；LSTM；注意力（Attention）机制
中图分类号：TP389.1 **文献标志码**：A **文章编号**：1672-6510（2020）06-0072-09

Office Staff Behavior Recognition Based on ResNet-LSTM with Attention Mechanism

ZHANG Chuanlei，WU Dashuo，XIANG Qihuai，CHEN Jia，LIU Lixin
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Aiming at the problem of confusion and high computational cost of feature extraction in time domain and space domain in human behavior video analysis，a ResNet-LSTM-Attention network model is designed. The network extracts features in space domain and time domain separately. For spatial dimension，frame images are put into ResNet model to extract spatial dimension features. As to time dimension，multi-frame superimposed spatial dimension features are used as the input of the LSTM and Attention fusion model. Finally，the output of softmax is the recognition result. The network model is applied to the office staff behavior video dataset collected in the laboratory for human behavior recognition experiments，and then compared with the ResNet-LSTM model without Attention mechanism and C3D network model commonly used in video analysis. The experimental results show that the new model not only has higher recognition accuracy，but also reduces the complexity of network parameters and operation costs.

Key words：behavior recognition；video analysis；residual network；LSTM；Attention mechanism

基于视频的人体行为识别技术作为计算机视觉领域研究热点之一，具有较高的科学研究价值和应用价值，包括对视频中图像序列自动进行人体行为检测、识别和理解等相关内容. 目前而言，关于人体行为识别的研究较多，但很少涉及办公领域. 通过对办公大厅内的监控视频进行分析，能够有效了解办公人员工作状态、工作习惯等，从而可以制定合理规章制度，督促人员合理安排工作时间，提高工作效率和服务质量. 因此，进行基于视频分析的办公人员行为识别研究具有重要的应用价值.

收稿日期：2019–10–08；**修回日期**：2020–01–12
作者简介：张传雷（1973—），男，山东省淄博人，教授；**通信作者**：武大硕，硕士研究生，wudashuo@gmail.com

关于人体行为识别的研究最早开始于 19 世纪 70 年代左右，科学家在动物行为方面展开了机械学研究，但是由于当时计算机发展水平较低，计算资源有限，无法支持大量的科学计算，人体行为分析没有得到相应的重视[1]. 到了 20 世纪 90 年代，为了对战场以及日常民用视频监控等场景下的视频进行分析和理解，美国国防部高级研究计划局、麻省理工学院和卡内基梅隆大学等多所高校参与了视觉监控系统研究. 在法国，由国家信息与自动化研究所成立 WILLOW 小组，主要致力于研究分析人体行为的分类和复杂场景识别等，而其成立的 PRIMA 小组主要研究单个个体或者人群的行为识别. 欧盟也设立了 ADVISER 项目，致力于研究智能交通管理系统、人机交互和人体行为分析与理解等[1]. 国内也有很多高校和研究机构进行人体行为识别的相关研究，包括清华大学、北京大学、中科院自动化模式识别国家重点实验室、北京航空航天大学等[2].

在深度学习应用到行为识别领域前，国内外研究学者对基于手工特征的行为识别方法进行了广泛研究. Bobick 等[3]提出基于轮廓剪影进行特征提取，通过轮廓剪影建立运动能量图来描述人体步态动作. 这种方法在简单背景下的描述能力较强，但在背景相对较复杂的情况下效果不佳. Peng 等[4]提出基于时间序列引入对背景光流和轨迹的消除方法——iDT 方法，使特征更加集中于人体运动的描述. iDT 方法是深度学习进入该领域前效果、稳定性、可靠性最高的方法，不过算法复杂度很高. 传统的行为识别方法不具有普适性，基于深度学习从数据中自动学习特征的方法效果更优.

近年来，基于计算机深度学习模型的特征学习引起研究人员的广泛关注，基于深度学习模型的特征提取也成为一个重点的研究对象. 在传统的机器学习中，往往是通过传统算法提取特征，这样会使结果更偏向于局部特征的表现，忽略了全局特征，从而造成局部特征提取对缩放、角度变换等因素不敏感. 近几年，作为深度学习模型之一的卷积神经网络（convolution neural network，CNN）在图像识别、语音识别、视频处理等领域取得了巨大成功，基于卷积神经网络的特征提取，可直接以图像矩阵作为模型的输入，避免了像传统机器学习那样前期对图像数据的各种复杂的预处理，实现了监督式的学习，由局部到全局、由低级到高级的特征提取. CNN 一般由输入层、输出层和多个隐藏层组成，隐藏层一般包括卷积层、池化层、激活层和全连层.

CNN 在视频中应用的一个方法是对每一帧用 CNN 进行识别，但这种方法只考虑到了空间上的视觉效果，没有考虑到行为运动是一个序列，在时间维度上还有关联，连续帧之间有一定耦合，是相互关联的[5]. 因此，Simonyan 等[6]提出了 Two-Stream 结构的 CNN，此网络不仅包括空间维度还包括时间维度，空间流处理静止的图像帧，得到形状特征；而时间流处理连续数帧稠密光流[7]，可以提取动作信息，利用多任务训练的方法把这两个数据集结合起来，但是两个流都是 2D 卷积操作，不能很好地提取时间特征.

针对 2D 卷积不能很好地提取时间特征，Tran 等[8]提出了一个比较经典的 C3D 网络来提取视频的空间特征和时域特征. 这是首次提出 3D 卷积网络，让 3D 卷积网络逐渐成为研究热点. 相比于 2D 卷积网络，3D 卷积网络能够更好地提取空间特征和时间特征，而且只需要配合简单的分类器就能有很好的表现. 其使用 $3\times3\times3$ 的卷积核在实验中比其他几个结构都要好，得出的结构特征通过线性分类器后，几乎可以达到当时最好的精度. 3D 卷积虽然能很好拟合时间域和空间域上的特征，但在时空两个维度同时反向传播进行权重修正时，也很容易造成两个维度上一定程度的特征提取混乱. 同时，3D 卷积网络的网络结构的参量和运算成本相对于 2D 卷积网络而言要大的多.

Carreira 等[9]提出了 I3D 网络，I3D 用于图像分类的 2D 卷积网络变形成可以提取时空特征的特征提取器，弥补了 3D 卷积网络参数多以及需要从零开始训练的不足，相较于 C3D 网络有显著提升. Donahue 等[10]提出长时循环卷积神经网络（longterm recurrent convolutional network，LRCN），其将 CNN 与 LSTM 相结合，通过 CNN 提取单帧图像的卷积特征并将其按时间顺序输入 LSTM 中，最终得到视频数据的行为特征.

本文针对视频分析中空间和时间两个维度的特征，提出一种卷积神经网络、循环神经网络和注意力模型的融合模型（ResNet-LSTM-Attention）. 对于空间维度，将单帧图像输入 ResNet 模型，提取空间维度特征；对于时间维度，将多帧叠加后的空间维度特征作为输入，输入到循环神经网络（LSTM）和注意力（Attention）模型的融合网络模型；然后将 ResNet-LSTM-Attention 模型的输出经过 Softmax 输出作为结果，得到一个多模型融合的视频人体行为识别的网

络模型. 最后将本文网络模型结构应用到办公领域人员行为视频分析. 本文提出网络模型优势在于将空间域和时间域的特征分开提取. 首先在静止的图片上提取特征，随后在时间序列上分别进行拟合. 该模型的两层神经网络相互独立，训练过程分开进行，所以提取时间域的 LSTM 的反向传播不会贯穿到 ResNet，从而一定程度上避免造成时间域和空间域上特征提取的混淆.

1　网络模型理论分析

1.1　ResNet 神经网络模型

深度卷积神经网络(deep convolutional neural networks，DCNN)在数据分类领域应用广泛并且取得了巨大的突破，例如语音和文字、视频和图像方面，这是因为 DCNN 具有 3 个重要特征：局部区域感知、时空域上采样和权重共享. 但是深度卷积神经网络也存在 3 个问题：(1)常规的网络并不是随着网络层数增加，堆叠效果会更好；(2)网络层数越深，会出现梯度消失问题，使得训练效果不会很好；(3)层数较浅的网络通常不会使识别效果明显提升[11].

为了解决上述 3 个问题，He 等[12]提出了残差网络 ResNet，引入了残差块(residual block)构建深层网络，残差块结构如图 1 所示. 其中 x 为输入，$H(x)$ 为输出，$F(x)$ 为残差映射函数，weight layer 为卷积层.

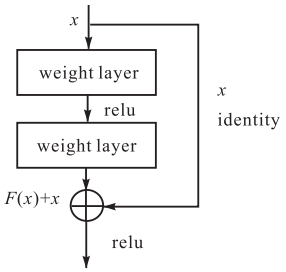


图 1　ResNet 残差块结构

Fig. 1　Residual block structure of ResNet

构建深层网络的基本约束条件，使堆叠后的网络模型误差应比基础的浅层模型更低，因此在实际中采用恒等映射的方法构建深层模型，即用 $H(x)=x$ 作为最佳解映射. 当层数较深时，模型难以直接拟合实际映射 $H(x)$，因此 ResNet 引入了“shortcut”快捷连接，这就使问题转换为拟合残差映射 $F(x)$，此时实际映射 $H(x)$ 表示为 $H(x)=F(x)+x$. 当 $F(x)=0$ 时，就构成了一个恒等映射 $H(x)=x$，模型只需最小化残差函数 $F(x)=H(x)-x$ 来逼近实际映射以解决网络层堆叠的性能退化问题[13].

现假设有共计 L 层残差块连接，$x^{(l)}$ 表示第 l 个残差块的输入，$x^{(l+1)}$ 表示该残差块的输出，也是第 $l+1$ 个残差块的输入. 可得第 l 个残差块的输出为

$$x^{(l+1)}=x^{(l)}+\sum_{i=1}^{L-1}F(x^{(i)}+W^{(i)}) \tag{1}$$

由式(1)可见，每层残差网络都在累积上层的残差特征，保证了 $l+1$ 层始终比 l 层拥有更多的特征信息，第 L 层始终拥有最多信息. 在反向传播过程中，根据链式求导法则，误差损失项 loss 对于网络前端的第 l 个残差块的梯度计算式为

$$\frac{\partial \text{loss}}{\partial x^{(l)}}=\frac{\partial \text{loss}}{\partial x^{(L)}}\frac{\partial x^{(L)}}{\partial x^{(l)}}=\frac{\partial \text{loss}}{\partial x^{(L)}}\left(1+\frac{\partial}{\partial x^{(l)}}\sum_{i=l}^{L-1}F(x^{(i)}+W^{(i)})\right) \tag{2}$$

由式(2)中的因式分解项 $\frac{\partial \text{loss}}{\partial x^{(L)}}$ 可知，残差网络最深一层 L 的信息可以直接反向传播到任意较浅的一层 l 中去；又因项 $\frac{\partial}{\partial x^{(l)}}\sum_{i=l}^{L-1}F$ 的值不可能一直为−1，因此即使每层的权重再小，也不会出现梯度消失问题.

本文针对视频分析中空间维度的特征，采用 ResNet 模型.

1.2　LSTM 神经网络模型

在深度学习中能良好表达时序的网络结构是循环神经网络(recurrent neural network，RNN)，其中表现最优的是 LSTM. 由于 LSTM 是对序列进行操作，多层的 LSTM 堆叠可使输入的抽象级别增加，当时间增大即可分块观察，或在不同的时间尺度上表示问题，使得网络能提取出更加抽象的特征，所以本文通过堆叠多层 LSTM 进行时间域的特征提取. 本文所研究的办公人员视频分析问题是典型的时序问题，即某一个时刻的值受前一个时刻或几个时刻的影响[14]，因此选择 LSTM 模型.

LSTM 属于时序卷积神经网络，是由循环神经网络衍生而来的，通过引入门函数，可以挖掘时间序列中相对较长间隔和延迟等的时序变化规律[7]. 图 2 为 LSTM 内部结构. 图中：x_t 为第 t 个输入序列元素值；c 为细胞状态或称为记忆单元，控制信息的传递，也是网络的核心；i 为输入门，它决定了当前 x_t 保留多少信息给当前状态 c_t；f 为遗忘门，它决定保存多少前一时刻的细胞状态 c_{t-1} 至当前的 c_t；o 为输出门，它决定 c_t 传递多少至当前状态的输出 h_t；h_{t-1} 指代在 $t-1$ 时刻的隐藏层状态[15].

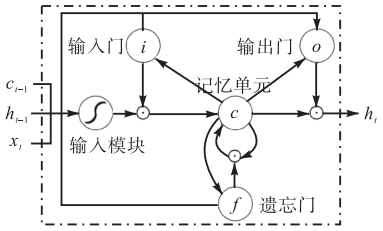


图 2　LSTM 内部结构图

Fig. 2　LSTM internal structure diagram

上述过程对应式(3)—式(8).

$$i_t=\sigma(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+b_i) \tag{3}$$

$$f_t=\sigma(\boldsymbol{W}_{xf}x_t+\boldsymbol{W}_{hf}h_{t-1}+b_f) \tag{4}$$

$$\mathrm{o}_t=\sigma(\boldsymbol{W}_{\mathrm{xo}}x_t+\boldsymbol{W}_{ho}h_{t-1}+b_o) \tag{5}$$

输入门 i_t、遗忘门 f_t 和输出门 o_t 的结果均为当前输入序列 x_t 和前一状态输出 h_{t-1} 乘以相对应权重加上对应偏移量，最后经过 sigmoid 激活函数所得. 而当前时刻单元的即时状态 $\overline{c_t}$ 则使用 tanh 激活函数激活，见式(6).

$$\overline{c_t}=\tanh(\boldsymbol{W}_{xc}x_t+\boldsymbol{W}_{hc}h_{t-1}+b_c) \tag{6}$$

而新的单元状态 c_t 则由当前记忆 $\overline{c_t}$ 和长期记忆 c_{t-1} 结合而成，按式(7)计算.

$$c_t=f_t\cdot c_{t-1}+i_t\cdot\overline{c}_t \tag{7}$$

则 LSTM 单元的输出 h_t 的计算式为

$$h_t=o_t\cdot\tanh(c_t) \tag{8}$$

上述公式中，$\boldsymbol{W}_{xi}$、$\boldsymbol{W}_{xf}$、$\boldsymbol{W}_{xo}$、$\boldsymbol{W}_{xc}$ 分别是输入层到输入门、遗忘门、输出门与细胞状态的权重向量；而 $\boldsymbol{W}_{hi}$、$\boldsymbol{W}_{ho}$、$\boldsymbol{W}_{hf}$、$\boldsymbol{W}_{hc}$ 分别是隐藏层到输入门、输出门、遗忘门与细胞状态的权重向量；b_i、b_o、b_f、b_c 分别是遗忘门、输入门、输出门与细胞状态的偏移量；$\sigma(\cdot)$ 为 sigmoid 激活函数；tanh 为双曲正切激活函数；$\cdot$ 表示向量元素乘.

图 3 为 LSTM 分类模型.

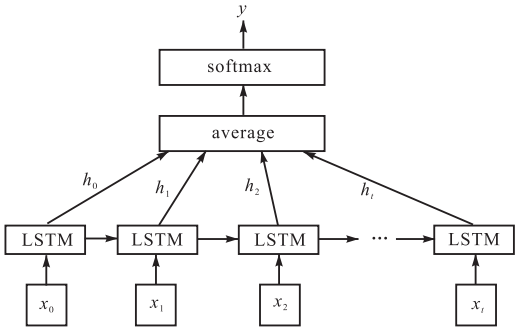


图 3　LSTM 分类模型

Fig. 3　LSTM classification model

图 3 中的输入层是对应的视频帧特征向量，在输入层上层是正向的 LSTM 层，由一系列的 LSTM 单元构成. 再将全部时刻的 LSTM 输出进行加权平均操作后的结果作为上层的表示. 最后通过 softmax 层，进行全连接的操作，最终可以得到预测结果的类别 y[16].

1.3　Attention 机制

Attention 机制即注意力机制，通常被运用在图像处理和自然语言处理领域. 学者们提出了不同种类的注意力机制，识别效果比较明显. 针对办公人员行为识别问题，本文对 LSTM 模型引入了注意力机制，它能对输入序列提取特征信息，寻找特征信息之间的时序内在联系，并通过加权平均方式给出识别结果，从而提高模型的识别准确度. 对于一系列权重参数，Attention 机制主旨思想是从序列中学习到每一个元素的重要程度，并按其重要程度将元素合并. 加入 Attention 机制可以使模型的性能得到显著提升；另外，使用 Attention 机制也可以观察到输入序列中的信息是怎样影响最后的输出序列，有助于更好地理解模型的内部运作机制，更便于对一些预设的输入与输出进行参数调试. 因此，在模型构建中本文在 LSTM 后接入一层 Attention 网络进行时序特征提取. 图 4 为 LSTM-Attention 分类模型.

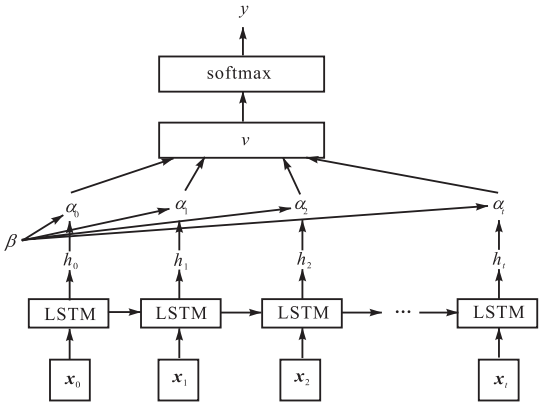


图 4　LSTM-Attention 分类模型

Fig. 4　LSTM-Attention classification model

图 4 中输入序列 $\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t$ 表示视频帧空间特征的向量，将输入依次传入到 LSTM 单元后，得到对应隐藏层的输出 $h_0,h_1,h_2,\cdots,h_t$. 同时，在隐藏层中引入 Attention 机制，计算每个输入分配的注意力概率分布值 $\alpha_0,\alpha_1,\alpha_2,\cdots,\alpha_t$，其思想是计算该时刻的隐藏层输出与视频帧空间特征向量的匹配得分占总体得分的比重[17]，设 h_i 为第 i 个时刻隐藏层的输出状态，$\overline{\boldsymbol{h}}$ 为比视频帧高一级的特征表示向量. 将 $\overline{\boldsymbol{h}}$ 进行随机初始化，作为一个参数在训练过程中逐步更新，$\alpha_i,i\in[0,t]$ 的计算式[18]为

$$\alpha_i = \frac{\exp(\beta_i)}{\sum_{j=1}^{n}\exp(\beta_j)} \tag{9}$$

其中：β_i 表示第 i 个隐藏层输出 h_i 在视频帧表示向量 $\bar{\boldsymbol{h}}$ 中所占的分值，β_i 越大，说明这个时刻的输入在整体中的注意力越大，它的计算公式为

$$\beta_i = \boldsymbol{V}^{\mathrm{T}}\tanh(\boldsymbol{W}\bar{\boldsymbol{h}} + \boldsymbol{U}h_i + b) \tag{10}$$

式中：$\boldsymbol{V}$、$\boldsymbol{W}$、$\boldsymbol{U}$ 为权值矩阵；b 为偏置量；tanh 为非线性激活函数.

各个时刻的注意力概率分布值经计算得出后，再计算包含特征信息的特征向量 $\boldsymbol{\varepsilon}$，公式为

$$\boldsymbol{\varepsilon} = \sum_{j=1}^{t}\alpha_j h_j \tag{11}$$

最后，经 softmax 分类函数后可得预测类别 y，计算式为

$$y = \mathrm{softmax}(\boldsymbol{W}_v v + b_v) \tag{12}$$

本文训练模型的迭代方法采用梯度下降法，通过计算损失函数的梯度并更新模型的参数，最终到达收敛. 为了使目标函数更加平稳地收敛，同时也为了提高算法的效率，每次只取小批量样本进行训练. 模型使用的损失函数为交叉熵，计算式为

$$H_{y'}(y) = -\sum_i y_i' \log y_i \tag{13}$$

其中 y_i' 是实际类别标签值，y_i 是经 softmax 分类函数计算后的预测类别标签值.

2　数据处理及模型设计

2.1　数据获取及预处理

2.1.1　数据获取

本文所用的包括训练集、测试集和验证集数据均是实验室自行采集. 获取数据的步骤：(1)将所有动作录制成视频；(2)将视频每 10 帧抽 1 帧，即每秒抽取约 3 帧图片；(3)将图片中主要表现的人体行为部分进行裁剪.

所采集的视频数据共分为 8 类，分别是打电话、吃东西、离岗、玩手机、睡觉、抽烟、工作和交流，数据集示例图片如图 5 所示.

为了充分利用计算资源，本文将所有数据集做成了队列的形式，分批读入内存缓冲区，训练数据依次从缓冲区里读取，使用的方法为 TFRecord，它是 TensorFlow 提供的一种数据存储办法. TFRecord 理论上可以保存任何格式的信息，可以将任何类型数据转化为 Tensorflow 所支持的格式，这种方法可以让数据集和网络模型更容易相互适应匹配，此外利用 TFRecord 可以很方便实现队列.

图 5　数据集图片

Fig. 5　Dataset image

2.1.2　数据增强及预处理

首先对图片进行了分类，将截取下来的图片进行了手工标注，标注为同一动作的图片序列归于同一文件夹中. 随后，对图片进行了分组. 将现有数据集分为两组，其中一组从中抽取部分有代表性的关键帧进行 CNN 网络训练，训练集每个动作抽取 1 000 张图片，验证集每个动作抽取 200 张图片，共计 9 600 张图片，并使用 OpenCV 进行图片的预处理以固定图片大小为 283×240，在训练的时候进行随机裁剪和图片增强. 另一组数据集每 16 帧为一组，每个动作分出若干组序列帧，将图片统一大小为 224×224. 由于本组图片直接用于已经训练好的 CNN 模型提取概率特征，随后进一步提取时间特征，所以不需要随机裁剪. 这样共有 7 066 组训练集，1 347 组验证集，共计 134 608 张图片.

此外针对 C3D 模型的训练，将图片每 16 帧分为一组，每张图片裁剪为 171×128，在训练的时候进行随机裁剪和增强，同样也得到 7 066 组训练集与 1 347 组验证集.

最后，对数据集进行增强与归一化处理. 在实际的训练过程中，数据集偏少，所以使用在线增强数据集的方式来扩充训练数据，即应用模型进行训练时，首先获得一个 batch 数据，然后对这个 batch 的数据进行随机增强，同时通过 GPU 优化计算. 此外，由于图像数据是 0 ~ 255 的 uint 数据，本文对图像进行归一化处理，使图像数据转化为介于 0 ~ 1 之间分布的数据，若原始图像数据为 x，则本文使用最常用的最大最小值归一化方法按式(14)计算.

$$norm = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

其中：x_i 表示图像像素点的值；x_{max} 和 x_{min} 分别表示图像像素的最大值和最小值. 通过归一化的方法，可以有效防止仿射变换的影响，减小集合变换的影响，同时加快梯度下降求最优解的速度[19].

2.2　深度神经网络模型设计

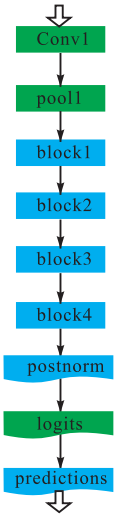
本文设计的 ResNet-LSTM-Attention 网络模型的网络结构共两层，分别为图像特征提取层和时序特征提取层，图像特征提取层提取图片在二维空间上的特征，时序特征提取层提取图像序列之间的时序特征.

2.2.1　图像特征提取层

图像特征提取层本文使用的残差网络（ResNet），该网络结构能很好地解决 CNN 增加深度会造成梯度弥散或者梯度爆炸的问题. 本文在网络模型构建中使用的为 50 层的 ResNet 网络，其结构参数和数据流图如图 6 所示.

layer name	output size	50-layer
conv1	112 × 112	7 × 7, 64, stride 2
conv2_x	56 × 56	3 × 3 max pool, stride 2 [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, 1 024] ×6
conv5_x	7 × 7	[1×1, 512; 3×3, 512; 1×1, 2 048] ×3
	1 × 1	average pool, 1000-d fc, softmax
FLOPs		3.8×10^9

（a）结构参数



（b）数据流图

图 6　残差网络结构参数和数据流图

Fig. 6　Residual network structure parameters and data flow diagram

网络分为 5 个隐藏层，分别为 conv1、conv2_x、conv3_x、conv4_x 和 conv5_x. ResNet-50 首先输入 7 × 7 × 64 的卷积，随后经过 3 + 4 + 6 + 3 = 16 个 building block，每个 block 为 3 层，即有 16 × 3 = 48 层，最后连接全连接层，所以共 1 + 48 + 1 = 50 层（这里仅仅指的是卷积层或者全连接层，激活层或池化层并没有计算在内）.

本文输入图片的大小为 224 × 224 × 3，首先经过第一个卷积核为 7 × 7，步长为 2 的卷积层，图片降维度到 112 × 112 × 64，然后经过一个核为 3 × 3，步长为 2 的最大池化层，之后依次进入 block1、block2、block3、block4 这 4 个残差块，每个残差块有 3 层卷积层，输出 7 × 7 × 2 048 的向量，随后连接上一层平均池化层，输出 1 × 1 × 2 048 的特征向量，最后连接一层全连接层，输出得分向量（未归一化的概率向量）. 由于本文一共进行 8 类动作的分类，所以最终图像特征提取层的输出为 8 个概率特征向量.

2.2.2　时序特征提取层

时序特征提取是在已有的概率特征向量序列上进行时域上的特征提取，包括输入层（in）、LSTM 层、Attention 层和输出层（out），下面将结合图 7 的 LSTM-Attention 数据流图逐层进行阐述.

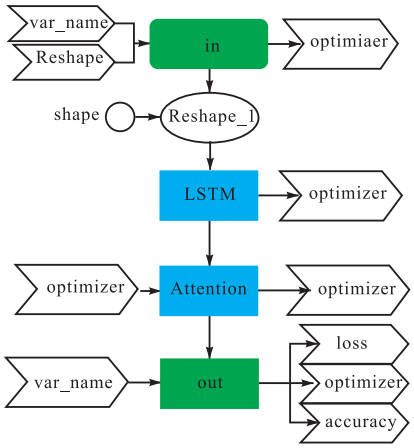


图 7　LSTM-Attention 数据流图

Fig. 7　LSTM-Attention data flow diagram

in 层的输入是图像特征提取层的输出，目的在于对图像概率特征向量进行放大处理. 随后为两层具有 128 个单元的 LSTM，第一层 LSTM 的输出作为第二层 LSTM 的输入，即 $x_t^2 = h_t^1$. LSTM 层后紧跟 Attention 网络，通过引入 Attention 对序列进行权重分配. 网络模型的最后为 out 层，Attention 层输出为加权后的得分向量，输出每个元素的维度为 1*128，最后再连接 out 层，对得分向量进行降维，最后的输出为 1*8 的未归一化的概率向量. 即最开始输入时序提取层的维度为 batch_size*16*8，至输出层输出的维度为 batch_size*8.

本文对于图像特征提取层和时序特征提取层的具体步骤总结如下：

（1）截取视频中动作的关键帧，训练出准确率较高的 ResNet 模型，从而使得每帧图片的行为类别的可能性体现在最后的得分向量中.

（2）整理序列帧，将序列中每一帧分别输入训练好的 ResNet 模型，得到 logits 序列，即未归一化的概

率序列.

(3) 对每帧的得分向量进行特征放大，随后进入时序提取层，通过连接输出层将 softmax 概率归一化.

3　实验与结果分析

将数据进行预处理后，对 ResNet-LSTM-Attention 模型与 C3D 模型进行实验结果分析和对比. 本文代码基于 TensorFlow 实现，运行环境：操作系统 Windows 10，Python 版本 Python3.6，TensorFlow 版本 Tensorflow-1.11.0，GPU 驱动为 CUDA9.0 与 CUDNN7.1.

实验步骤分为定义阶段、训练阶段和评估阶段. 其中：定义阶段包括对于模型结构、损失函数及优化器等的定义[20]，具体定义指标见表 1，其中 Dropout 参数取的是 0.8，L2 正则化 lambda 值取的是 0.005，最大 batch 值为 200 000，收敛阈值为 0.01，即当训练集损失低于 0.01 时视为完全收敛，并记录此时模型收敛时间与 batch 数. 训练阶段使用 3 个模型对相同的数据进行训练与测试，本文提出的 ResNet-LSTM-Attention 模型作为实验组，C3D 和没有 Attention 机制的 ResNet-LSTM 模型作为对照组.

表 1　模型定义指标

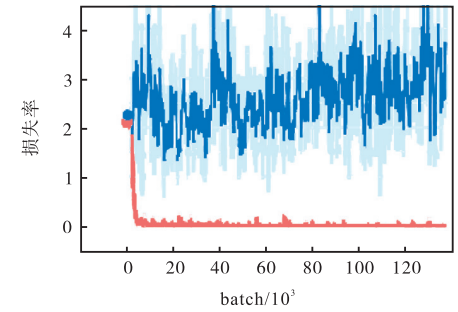
Tab. 1　Model parameters

类别	指标
优化器	Adam 优化器
防止过拟合	Dropout 和 L2 正则化方法
损失函数	交叉熵
学习率	指数衰减方式

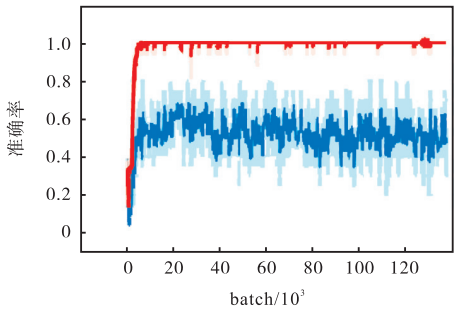
首先将数据训练 C3D 模型，C3D 模型损失率及准确率变化如图 8 所示，其中橙色的线代表训练操作，蓝色的线代表验证操作. 在 C3D 模型中，大约经过 3 000 个 batch 后模型开始收敛，准确率逐渐上升，损失逐渐下降. 经过 130 000 个 batch 后，训练集的损失收敛到 0.007 左右，准确率达到 1；验证集的损失收敛在 1 ~ 3，准确率达到 0.55 左右.

然后，使用 ResNet-LSTM 模型对同一数据集进行训练和验证. ResNet-LSTM 模型损失率和准确率变化如图 9 所示. 在大约 2 000 个 batch 后开始收敛，准确率逐渐上升，损失逐渐下降. 经过 200 000 个 batch 后，训练集的损失收敛到 0.002 左右，准确率达到 1；验证集的损失收敛在 1.56 左右，准确率达到 0.73 左右.

最后，将同样的数据验证本文提出的 ResNet-LSTM-Attention 模型，由于 Attention 机制的引入，将图像特征提取层和时序特征提取层分开来看. 图像特征提取层损失率和准确率变化如图 10 所示.



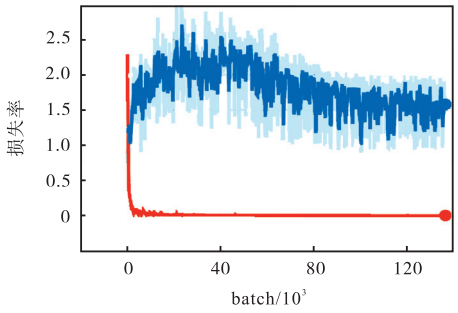
(a) 损失率变化



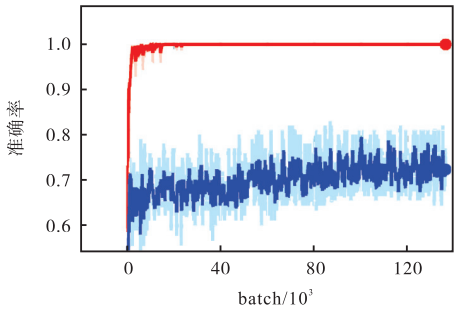
(b) 准确率变化

图 8　C3D 模型的损失率和准确率变化

Fig. 8　Loss change and accuracy change in C3D model



(a) 损失率变化

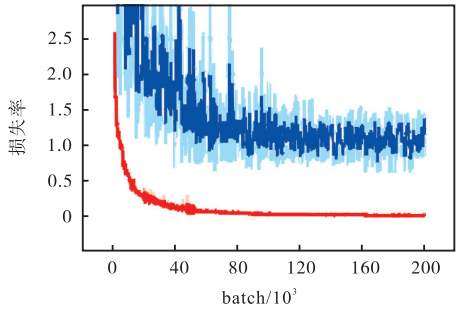


(b) 准确率变化

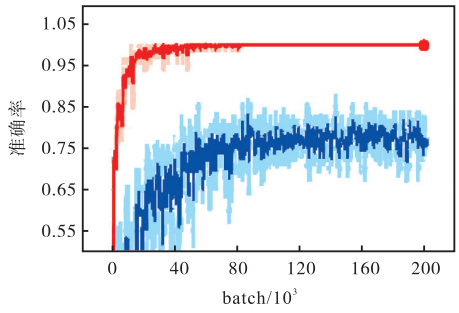
图 9　ResNet-LSTM 模型的损失率和准确率变化

Fig. 9　Loss change and accuracy change in ResNet-LSTM model

在图像特征提取层中，大约经过 2 000 个 batch 后模型开始收敛，准确率逐渐上升，损失逐渐下降. 经过全部 200 000 个 batch 后，训练集的损失收敛到 0.008 左右，准确率达到 1，即对训练集的分类全部正确；而对验证集的损失收敛到 1 左右，准确率达到 0.75 左右.



（a）损失率变化



（b）准确率变化

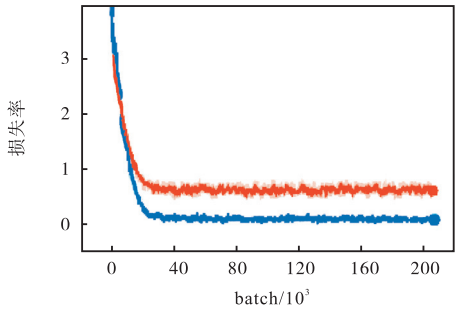
图 10　本文模型图像特征提取层损失率和准确率变化

Fig. 10　Loss change and accuracy change in image feature extraction layer of the new model

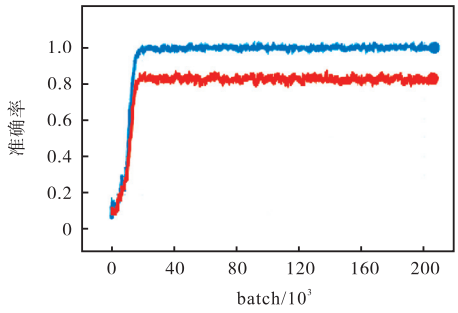
经过时序特征提取层后，损失率和准确率变化如图 11 所示. 在时序特征提取层，约经过 7 000 个 batch 后模型开始收敛，准确率逐渐上升，损失逐渐下降. 经过 200 000 个 batch，训练集的损失收敛到 0.09 左右，准确率达到 0.96 左右；对验证集的损失收敛到 1 左右，准确率达到 0.8 左右.

上述实验结果表明 ResNet-LSTM-Attention 的网络结构最终对验证集可以达到 0.8 左右的准确率，高于无 Attention 机制的模型 7 个百分点，并且远远高于经典的 C3D 模型，证明了本文提出方法的可行性. 无 Attention 机制的模型相较于本文提出模型，虽然在训练集的损失较低，精度较高，但验证集却全面落后，证明了其稍微出现过拟合现象，而本文模型由于 Attention 机制的加入，能够更好提取重点特征，鲁棒性加强，验证集精度为所有模型中最高，表现最好.

从计算速度方面看，C3D 用时 348 min，远高于 ResNet-LSTM 模型（294 min）和 ResNet-LSTM-Attention 模型（266 min），后两个模型虽同为 2D 卷积模型，但本文提出的具有注意力机制的模型用时比无注意力机制的模型少 28 min.



（a）损失率变化



（b）准确率变化

图 11　本文模型时序特征提取层损失率和准确率变化

Fig. 11　Loss change and accuracy change in time series feature extraction layer of the new model

4　结　语

本文提出了一种基于 ResNet-LSTM-Attention 网络结构的办公人员行为智能识别方法，并通过实验对比证明了使用 2D 卷积神经网络 ResNet 结合 LSTM 进行时序分类要比 3D 模型用时少，精度高；而注意力机制的加入使得模型鲁棒性增强，减少过拟合程度，并且训练用时和精度都有提升，论证了本文提出的方法具备一定的意义与价值. 在后续的研究中，本文将针对复杂环境下（如光线不良、有遮挡等）的视频数据进一步提升方法的性能，可以通过在多种复杂环境下采集数据集用以扩张训练集，同时通过对训练集进行图片增强的方式尝试解决泛化能力不足的问题.

参考文献：

[1] 李鸣，张鸿. 基于深度特征分析的双线性图像相似度匹配算法[J]. 计算机应用，2016，36（10）：2822–2825.

[2] 富倩. 人体行为识别研究[J]. 信息与电脑：理论版，2017（24）：146–147.

[3] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 23(3): 257–267.

[4] Peng X, Wang L, Cai Z, et al. Hybrid super vector with improved dense trajectories for action recognition[EB/OL]. [2019–10–08]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.718.5729&rep=rep1&type=pdf.

[5] 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.

[6] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Computer Science, 2014(1): 568–576.

[7] 李艳荻, 徐熙平. 基于空–时域特征决策级融合的人体行为识别算法[J]. 光学学报, 2018, 38(8): 306–319.

[8] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// IEEE. 2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 7410867.

[9] Carreira J, Zisserman A. Quo vadis, action recognition? A new nodel and the kinetics dataset[C] // IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 8099985.

[10] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677–691.

[11] 高磊, 范冰冰, 黄穗. 基于残差的改进卷积神经网络图像分类算法[J]. 计算机系统, 2019, 28(7): 139–144.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 7780459.

[13] 金余丰, 姚美常, 刘晓锋, 等. 残差网络和注意力机制相结合的滚动轴承故障诊断模型[J/OL]. 机械科学与技术: 1–7 [2019–09–11]. https://doi.org/10.13433/j.cnki.1003-8728. 20190166.

[14] 陈佳, 刘冬雪, 武大硕. 基于特征选取与 LSTM 模型的股指预测方法研究[J]. 计算机工程与应用, 2019, 55(6): 108–112.

[15] 李梅, 宁德军, 郭佳程. 基于注意力机制的 CNN-LSTM 模型及其应用[J]. 计算机工程与应用, 2019, 55(13): 20–27.

[16] 蓝雯飞, 徐蔚, 汪敦志, 等. 基于 LSTM-Attention 的中文新闻文本分类[J]. 中南民族大学学报: 自然科学版, 2018, 37(3): 129–133.

[17] 汪涛, 汪泓章, 夏懿, 等. 基于卷积神经网络与注意力模型的人体步态识别[J]. 传感技术学报, 2019, 37(7): 1027–1033.

[18] 陈煜平, 邱卫根. 基于 CNN/LSTM 和稀疏下采样的人体行为识别[J]. 计算机工程与设计, 2019, 40(5): 1445–1450.

[19] 刘嘉莹, 张孙杰. 融合视频时空域运动信息的 3D CNN 人体行为识别[J]. 电子测量技术, 2018, 41(7): 43–49.

[20] 王萍, 庞文浩. 基于视频分段的空时双通道卷积神经网络的行为识别[J]. 计算机应用, 2019, 39(7): 2081–2086.

责任编辑: 郎婧