文章编号：2095-2163(2023)09-0111-05 中图分类号：TP391.41 文献标志码：A

基于注意力机制的动态手势识别方法

黄 圣，茅 健
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：实时识别动态手势是一项艰巨的任务，因为系统永远无法知道手势在视频流中何时或从何处开始和结束。由于其各种应用，许多研究人员一直致力于基于视觉的手势识别。提出了一种基于 3D 卷积神经网络（3D-CNN）和长短期记忆（LSTM）网络相结合的深度学习框架，整个架构同时融合了注意力机制（CBAM）。所提出的架构从视频序列输入中提取时空信息，同时避免大量计算。3D-CNN 用于提取光谱和空间特征，然后将特征图像提供给注意力机制模块，在增强图像特定区域的表征能力的同时加强特征的表达，最后通过 LSTM 网络进行分类。实验结果表明，所提方法能很好地识别动态手势，识别率达到了 95.58%，验证了所提方法的有效性和可能性。

关键词：动态手势识别；3D 卷积神经网络；注意力机制；长短期记忆法；人机交互

Dynamic gesture recognition method based on attention mechanism

HUANG Sheng, MAO Jian
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 Recognizing dynamic gestures in real-time is a difficult task because the system can never know when or where the gestures begin and end in the video stream. Due to its various applications, many researchers have been working on vision-based gesture recognition. This paper proposes a deep learning framework based on the combination of 3D Convolutional Neural Network (3D-CNN) and Long Short-Term Memory (LSTM) network, and the whole architecture also incorporates the Attention Mechanism (CBAM). The proposed architecture extracts spatiotemporal information from video sequence input while avoiding computationally intensive. 3D-CNN is used to extract spectral and spatial features, and then provide the feature image to the attention mechanism module to enhance the representation ability of specific regions of the image while telling the model what to pay attention to, and finally classify it through the LSTM network. the experimental results show that the proposed method can recognize dynamic gestures well, and the recognition rate reaches 95.82%, which verifies the effectiveness of the proposed method. and possibility.

【Key words】 dynamic gesture recognition; 3D convolutional neural network; attention mechanism; long short-term memory method; human-computer interaction

0 引 言

人机交互系统是人与机器之间进行交流和信息传递的桥梁[1]。手势是人类有效表达自身想法的主要工具，其从简单到复杂的不同动作，使之能够与他人交流。随着科学技术的发展和人们对智能设备的应用需求的不断增加，通过机器识别肢体动作，成为研究热点之一[2]。

学习时空特征对于人类手势或动作识别的性能稳定至关重要。Li 等人[3]提出了一种具有注意力机制技术的三维卷积神经网络（3D-ConvNets），用于学习时空特征。该模型在特征的时空学习方面优于简单的 2D-CNN。Hakim 等人[4]提出使用 3D-CNN 模型和 LSTM 提取 23 个手势的时空特征，在分类阶段之后，将有限状态机（FSM）与 3D-CNN、LSTM 模型融合，以监督分类决策。从上述的研究中可以得出：时间信息和 LSTM 对于处理动态手势时获得准确的手势预测非常重要。因此，许多研究者开始利用混合模型来学习和进行动态手势识别任务[5]。

近年来，注意力机制作为深度学习领域的重大突破，通过计算特征信息的重要程度并分配权重来

作者简介：黄 圣（1996-），男，硕士研究生，主要研究方向：智能控制、模式识别；茅 健（1972-），男，博士，教授，硕士生导师，主要研究方向：航空装备检测与控制、智能机器人。
通讯作者：茅 健 Email：jmao@sues.edu.cn

收稿日期：2022-09-29

增强模型对重要特征的关注度。对此,本文提出了一种基于注意力机制的动态手势识别方法,结合3D-CNN、CBAM 和 LSTM 的混合模型使用,并在20BN-Jester 数据集上进行实验。

1 基础理论

1.1 3D 卷积神经网络(3D-CNN)

在 2D CNN 中,卷积层执行 2D 卷积,从前一层特征图上的局部邻域中提取特征,应用加性偏差,结果通过 sigmoid 函数传递。卷积应用于 2D 特征图,仅从空间维度计算特征;当应用于视频分析问题时,需要捕获编码在多个连续帧中的运动信息,为此在 CNN 的卷积阶段执行 3D 卷积,以计算空间和时间维度的特征[6]。将多个连续帧堆叠在一起形成立方体,将该立方体与 3D 内核进行 3D 卷积。

通过这种构造,卷积层中的特征图连接到前一层中的多个连续帧,从而捕获运动信息。形式上,第 i 层中第 j 个特征图上位置 (x,y,z) 的值由式(1)给出:

$$v_{ij}^{xyz}=\tanh\left(b_{ij}+\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ijm}^{pqr}v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

式中:tanh 是双曲正切函数,b_{ij} 是该特征图的偏差,m 是与当前特征图相连的第$(i-1)$ 层中特征图集上的索引数,w_{ijm}^{pqr} 是连接到前一层中第 m 个特征图内核的第 (p,q,r) 个值,R_i 是 3D 内核沿时间维度的大小,P_i 和 Q_i 分别是内核的高度和宽度。

在子采样层中,通过在前一层的特征图上对局部邻域进行池化,来降低特征图的分辨率,从而增强输入失真的不变性。可以通过以交替方式堆叠多层卷积和二次采样,来构建 CNN 架构,CNN 的参数(如偏差 b_{ij} 和核权重 w_{ijm}^{pqr})通常使用有监督或无监督方法来学习。

因为核权重会在整个立方体中复制,3D 卷积核只能从框架立方体中提取一种类型的特征。CNN 的一般设计原则是通过从同一组较低级别的特征图生成多种类型的特征,来增加后期层的特征图数量。与 2D 卷积情况类似,可以通过将具有不同内核的多个 3D 卷积,应用到前一层的相同位置来实现[7]。

1.2 长短时记忆网络(LSTM)

长短期记忆网络(LSTM)是对神经网络的扩展。LSTM 单元结构由输入门、输出门和遗忘门组成,其控制学习过程,内部结构如图 1 所示。这些门是在 sigmoid 函数的帮助下调整,以控制学习过程中的打开和关闭[8]。LSTM 中的长期记忆称为细胞状态,负责控制上一个 LSTM 单元格状态的信息,如果遗忘门输出状态为 0,则告诉单元门忘记信息,如果为 1,则告诉单元门将其保持在单元状态。

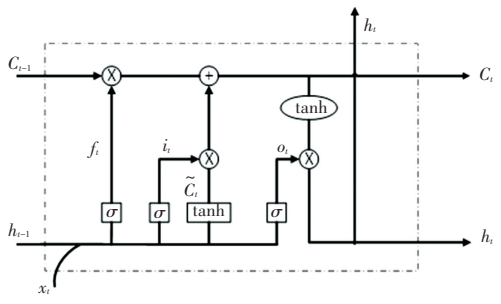

图 1 LSTM 网络单元

Fig. 1 LSTM network unit

LSTM 单元内的学习过程如式(2)~式(7):

$$f_t=\sigma(W_f[h_{t-1},\boldsymbol{x}_t]+b_f) \tag{2}$$

$$i_t=\sigma(W_i[h_{t-1},\boldsymbol{x}_t]+b_i) \tag{3}$$

$$\tilde{C}_t=\tanh(W_C[h_{t-1},\boldsymbol{x}_t]+b_C) \tag{4}$$

$$C_t=z_t*C_{t-1}+i_t*\tilde{C}_t \tag{5}$$

$$o_t=\sigma(W_o[h_{t-1},\boldsymbol{x}_t]+b_o) \tag{6}$$

$$\boldsymbol{h}_t=o_t+\tanh(C_t) \tag{7}$$

其中,i_t 是输入门;f_t 是遗忘门;o_t 是输出门;σ 为 sigmoid 激活函数;$\boldsymbol{x}_t$ 为 t 时刻的输入向量;w_x 为相应门的权重;b_x 为相应门的偏差;$\boldsymbol{h}_t$ 为 t 时刻的隐藏层状态向量;C_t 是 t 时刻 LSTM 单元的细胞状态。

遗忘门经由激活函数,输出一个 0~1 之间的数值,1 表示完全保留,0 表示完全舍弃。输入门通过 tanh 层创建候选状态,经由激活函数同样输出 0~1 之间的数值,决定候选状态 $\tilde{C}_t$ 需要存储多少信息。更新记忆单元将更新旧的细胞状态,将 C_{t-1} 更新为 C_t,遗忘掉由 f_t 确定的需要遗忘的信息,然后加上 $i_t*\tilde{C}_t$,确定新的记忆单元 C_t。输出门将内部状态的信息传递给外部状态 $\boldsymbol{h}_t$,经由激活函数层确定需要被传递出去的信息,将细胞状态通过 tanh 层进行处理并于输出门的输出相乘,最终外部状态会获取到输出门确定输出的那部分[9]。

1.3 CBAM 网络

在人类视觉大脑皮层中,使用注意力机制能够更快捷、高效地分析复杂场景信息,后来这种机制被研究人员引入到计算机视觉中来提高性能。注意力在告诉网络模型该注意是什么的同时也增强图像特定区域的表征能力[10]。Woo 等人[11]提出了一种结

合空间(spatial)和通道(channel)的注意力机制模块,被称为 CBAM,相较于单一的注意力机制,混合注意力机制显得更加全面。CBAM 模块能够针对一张特征图从通道和空间两个维度上产生注意力特征图信息,经过自适应修正产生最后的特征图。

如图 2 所示,通道注意力机制,通过特征内部之间的关系来获取最终的通道注意力值,特征图的每个通道都被视作一个特征检测器。通过同时采用平均池化和最大池化来压缩特征图的空间维度,实现更高效地计算通道注意力特征;将特征输入多层感知机(MLP)生成最终的通道注意力机制特征图 M_c。综上,通道注意力计算公式总结为式(8):

$$\begin{aligned} M_C(F) &= \sigma(\mathrm{MLP}(\mathrm{Avg}Pool(F)) + \\ &\quad \mathrm{MLP}(\mathrm{Max}Pool(F))) = \\ &\quad \sigma(W_1(W_0(F^c_{\mathrm{avg}})) + \\ &\quad W_1(W_0(F^c_{\mathrm{max}}))) \end{aligned}$$

其中,σ 为 sigmoid 函数;W_1、W_0 为 MLP 权重;F^c_{avg} 和 F^c_{max} 分别代表平均池化特征和最大池化特征。

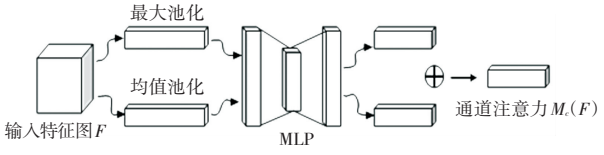

图 2　通道注意力模型

Fig. 2　Channel attention model

如图 3 所示,空间注意力机制是通过特征图空间内部的关系,来产生空间注意力特征图。为了计算空间注意力,首先在通道进行维度平均池化和最大池化,然后将其产生的特征图拼接起来,对拼接后的特征图中进行卷积操作,来产生空间注意力特征图 M_s。最终实现过程如式(9):

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([\mathrm{Avg}Pool(F);\mathrm{Max}Pool(F)])) = \\ &\quad \sigma(f^{7\times7}(F^c_{\mathrm{avg}};F^c_{\mathrm{max}})) \end{aligned} \tag{9}$$

其中,σ 为 sigmoid 函数,$f^{7\times7}$ 为 7×7 大小的卷积核。

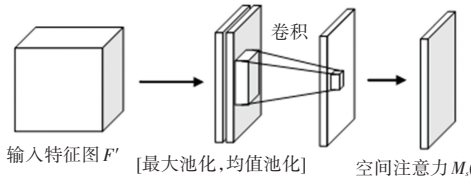

图 3　空间注意力模型

Fig. 3　Spatial attention model

CBAM 注意力机制模块结构如图 4 所示,其完整计算过程可以概括为如下公式:

$$F' = M_C(F) \otimes F \tag{10}$$

$$F'' = M_C(F') \otimes F' \tag{11}$$

其中,F 为输入特征图,F'' 为输出特征图。

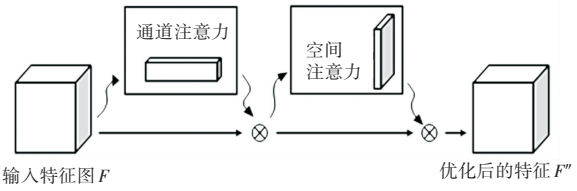

图 4　CBAM 完整结构

Fig. 4　CBAM complete structure

2　基于注意力机制的动态手势识别模型

2.1　模型提出

本文构建的是动态手势识别方法,数据集均为动态手势视频帧图像所组成的集合。由于数据采集环境不受限制,视频帧图像可能存在环境背景复杂、光线强弱等方面的问题,因此对模型的特征提取能力和抗干扰能力要求较高。传统卷积神经网络(CNN)模型虽然拥有较强的深层特征提取能力,但是空间信息特征提取能力不足,同时无法捕捉时间序列信息的前后关系。学习空间和时间特征的结合是动态手势分类的必要要求。为了实现这一点,本研究使用了 5 层 3D-CNN 模型,其可以通过保留视频帧的空间信息来提取时间特征。但是,仅仅使用 3D-CNN 模型进行动态手势识别还不足以从视频数据中学习长期的时空信息。LSTM 作为 RNN 的改进体,使用了一种特定的学习机制,明确了信息中需要被记住、需要被更新以及需要被注意的那些部分,以一种非常精准的方式来传递记忆,有助于在更长的时间内追踪信息。基于此,本文将 3D-CNN 与 LSTM 结合,使模型可以从空间和时间两个维度提取特征,使提取到的特征更加全面并且更具有代表性。

此外,在 3D-CNN 层后加入 CBAM,这种融合网络不会影响信息传输,同时模型可以自动学习得到图像的空间特征和通道特征的重要程度,根据重要程度来增强有用特征,自适应校准特征图像的空间和通道信息。相比于未添加注意力机制模块的网络模型,添加 CBAM 注意力机制模块对整体的网络结构影响不大,同时网络可以学习图像中更加重要的空间特征和通道特征。

2.2　模型结构

基于注意力机制的动态手势识别模型主要由 3D-CNN 层、注意力层、LSTM 层、Dropout 层以及

Softmax 层组成,模型结构如图 5 所示。

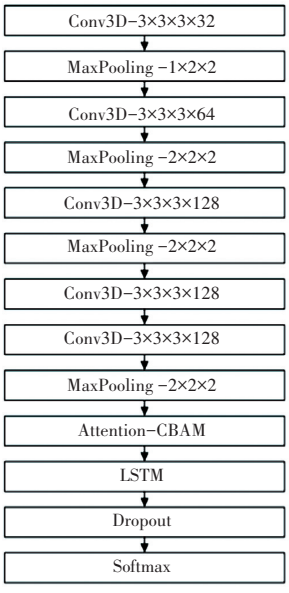

图 5 模型框架

Fig. 5 Model Framework

其中,3D-CNN 层由 5 个三维卷积层、4 个最大池化层组成,通过视频帧图像提取时空特征。每个 3 维卷积核大小均为 3×3×3,考虑到更好的保留时间细节,将第一卷积层和第一池化层的步幅和池化大小设置为 1×2×2,其余各层的步幅和池化大小设置为 2×2×2。特征图分别设置有 32、64、128 3 种不同的过滤深度。池化方式采用最大池化,用于保留主要特征信息。

将 3D-CNN 层提取到的特征图直接输入注意力层,CBAM 模块会根据输入的特征图,序列化的生成通道注意力机制特征图和空间注意力机制特征图,两种特征图信息与原特征图相乘进行自适应修正,产生最后的特征图。

将最终提取到的特征输入 LSTM 层,获取序列特征间的长期依赖关系。在 LSTM 层后添加一个值为 0.5 的 Dropout 层,保证输出的稀疏性,然后使用 Softmax 函数计算概率结果,实现动态手势的分类识别。

3 实验结果与分析

3.1 运行环境

本次实验环境的硬件配置为 Intel Core i7-11800H CPU,显卡为 NVIDIA RTX 3070。软件环境为 64 位 Ubuntu 20.04 操作系统,深度学习框架 PyTorch,Python 版本为 3.8.10。

3.2 数据预处理

实验使用 20BN-Jester 大规模真实数据集,该数据集由 1 376 个不同的参与者在不同的约束环境中生成。其中包含约 148 092 个 3 秒长的短视频片段,每个视频至少由 27 帧视频图像组成。由于时间及内存资源限制,这项工作仅使用了 27 个手势中的 12 个。在原始数据集中,视频序列具有不同的长度,从 27 帧到 46 帧不等。对于数据预处理,首先统一所有视频帧数,将每个视频片段统一为 30 帧视频图像来训练模型。对于每帧视频图像均调整为 112×112 像素。整个数据集中包含 12 个类共计 6 000个样本,每个类有 500 个样本。数据集按照 8∶2分为训练集和验证集,其中 80%为训练集和 20%为验证集。

3.3 实验结果讨论

为了证明本文方法的有效性,实验对比了 LSTM 网络、3D-CNN 网络、3D-CNN-LSTM 混合网络在同一数据集上的识别效果,各个方法的输入均为经过预处理后的数据集。LSTM 方法其模型主要由 3 个 LSTM 层、1 个全连接层和 Softmax 层组成,LSTM 单元数为 128,全连接层节点数为 64;3D-CNN 网络模型包含 5 个三维卷积层、4 个最大池化层、一个全连接层和 Softmax 层;3D-CNN-LSTM 混合网络模型包含 5 个三维卷积层、4 个最大池化层、1 个 LSTM 层、1 个 Dropout 层和 Softmax 层,卷积核尺寸为 3×3×3,LSTM 单元数为 128。

测试集数据包含 12 个动态手势类,每个类别分别包含 100 个文件。表 1 显示了使用 20BN-Jester 数据集中 12 个类在准确率方面对提出的混合模型与其它模型的比较结果。

表 1 数据训练集和测试集实验结果

Tab. 1 Experimental results of data training and testing sets

方法	训练集		测试集	
	准确率/%	损失	准确率/%	损失
3D-CNN	96.45	0.133 7	88.08	0.462 8
LSTM	94.02	0.281 6	88.94	0.342 7
3D-CNN +LSTM	98.58	0.032 27	91.22	0.378 9
3D-CNN+CBAM +LSTM	97.28	0.080 61	95.58	0.165 6

其中,本文提出模型对于动态手势实现了 95.58%的验证准确率,相比较于不包含注意力机制模块的模型,在准确率方向提高了 4.36%;相比较于单一模型的动态手势识别方法,准确率都有明显提升,该模型在取自 20BN-jester 数据集中 12 个类上产生了良好的结果。

如图 6、图 7 所示,从模型精度和模型损失曲线来看,由于模型是从头开始训练的,因此本次设置了

100 个 epoch 来达到所需要的损失。对于前 5 个 epoch，精度上升明显，损失非常高。后来，经过 10 个 epoch，模型达到了较高的准确率。经过 100 个 epoch 的训练，模型的验证准确率达到 95.58%且损失达到 0.165 6。

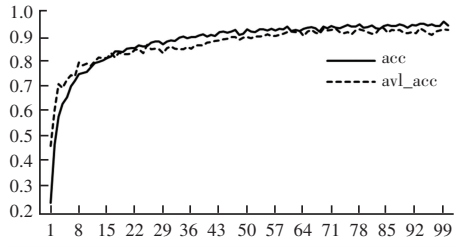

图 6　模型准确率

Fig. 6　Model accuracy

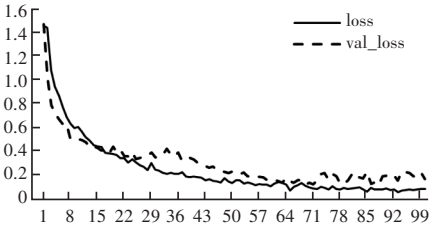

图 7　模型 loss

Fig. 7　Model loss

在总共 1 200 个视频剪辑中，有 100 个被归类为“向左滑动”手势。实际上，有 98 个视频片段属于向左滑动类，因此模型正确预测了 98 个片段，但有 2 个视频片段被预测为其他类别，因此该类别的识别准确率为 98%。所有类别中相对简单的手势如“竖起大拇指”手势，其识别准确率为 100%。同样对于所有剩余的类，分类结果显示在如图 8 所示的混淆矩阵中。

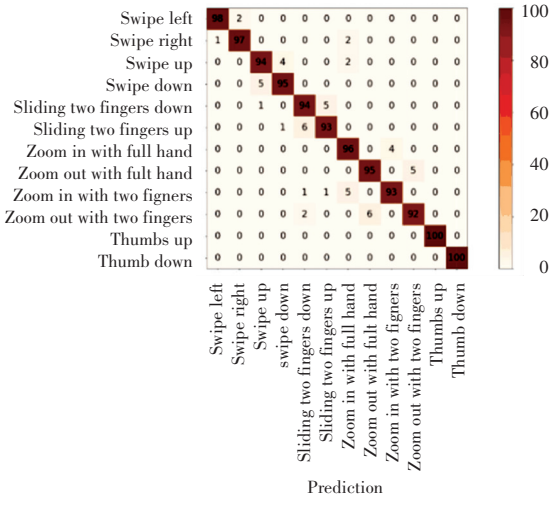

图 8　混淆矩阵

Fig. 8　Chaotic matrix

4　结束语

本文提出了一种新的深度学习模型，该模型可以学习视频流中动态手势序列的时空特征。该模型由 3D-CNN 网络、LSTM 网络和注意力机制网络组成，该网络在复杂的背景和照明条件下学习所有视频帧的空间和时间特征。在模型中，动态手势数据的特征由 3 维卷积神经网络（3D-CNN）自动提取；使用 CBAM 注意力机制网络增强特征关注度；使用长短期记忆（LSTM）网络来学习时间序列数据的相关优势；最后采用 SoftMax 分类器对动态手势进行分类。

经在 20BN-Jester 数据集的一个子集上进行训练，与单一模型和不包含注意力机制的混合模型相比，所提出的组合模型提供了更好的结果，动态手势识别性能更好。为了实现该算法的实际应用，后续工作会对算法的效率进行分析和提高。

参考文献

[1] WANG T, LI Y, HU J, et al. A survey on vision-based hand gesture recognition [C]//Smart Multimedia: First International Conference, ICSM 2018, Toulon, France, August 24 - 26, 2018, Revised Selected Papers. Cham: Springer International Publishing, 2018: 219-231.

[2] FANG L, FU M, SUN S, et al. Overview of face recognition methods[C]//Signal and Information Processing, Networking and Computers: Proceedings of the 5th International Conference on Signal and Information Processing, Networking and Computers (ICSINC). Springer Singapore, 2019: 22-31.

[3] JUN, LI, XIANG LONG, et al. Spatio-temporal deformable 3D ConvNets with attention for action recognition-Science Direct[J]. Pattern recognition, 2020, 98: 107037.

[4] Hakim N L, Shih T K, Kasthuri Arachchi S P, et al. Dynamic hand gesture recognition using 3DCNN and LSTM with FSM context-aware model[J]. Sensors, 2019, 19(24): 5429.

[5] WAN J, LI S Z, ZHAO Y, et al. ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2016:56-64.

[6] 梁正友，何景琳，孙宇.一种用于微表情自动识别的三维卷积神经网络进化方法[J].计算机科学，2020，47(8)：227-232.

[7] 佘海龙，解山娟，邹静洁.标准分数降维的 3D-CNN 高光谱遥感图像分类[J].计算机工程与应用，2021，57(4)：169-175.

[8] 谷学静，周自朋，郭宇承，等.基于 CNN-LSTM 混合模型的动态手势识别方法[J].计算机应用与软件，2021，38(11)：205-209.

[9] 麻文刚，张亚东，郭进.基于 LSTM 与改进残差网络优化的异常流量检测方法[J].通信学报，2021，42(5)：23-40.

[10] 王粉花，张强，黄超，等.融合双流三维卷积和注意力机制的动态手势识别[J].电子与信息学报，2021，43(5)：1389-1396.

[11] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]//European Conference on Computer Vision. Springer, Cham, 2018:3-19.