光电技术及应用

DOI：10.16818/j.issn1001－5868.2021081004

结合SoftPool的VGG19与CapsNet相级联的表情识别模型研究

杨　双，王敬东*，姜宜君，廖元晖

（南京航空航天大学 自动化学院，南京 210016）

摘　要： 表情识别作为计算机视觉领域的研究热点，在情感识别、人机交互、智能安防等领域有着重要的应用。文章针对VGG19在训练人脸表情数据集时由于全连接层参数量过大而易过拟合的问题，利用胶囊网络CapsNet对VGG19的全连接层进行替换，以实现VGG19与CapsNet相级联，从而改善训练时过拟合的问题，同时使得级联后的模型在RAF-DB数据集上的精度提高了5.28%。针对VGG19特征提取网络的MaxPool易丢失人脸特征图信息的问题，利用SoftPool对MaxPool进行替换，从而在最大程度上保留了人脸的细粒度特征。实验结果表明，改进后的模型在RAF-DB数据集上取得了84.21%的精度，在FER2013数据集上取得了73.16%的精度，表情识别效果更优。

关键词： 表情识别；VGG19；CapsNet；SoftPool

中图分类号： TP391.41　**文章编号：** 1001－5868(2021)06－0897－07

Research on Expression Recognition Model Cascading with VGG19 and CapsNet of SoftPool

YANG Shuang，WANG Jingdong，JIANG Yijun，LIAO Yuanhui

（College of Automation，Nanjing University of Aeronautics and Astronautics，Nanjing 210016，CHN）

Abstract： As a research hotspot in the field of computer vision，expression recognition plays an important role in emotion recognition，human-computer interaction，intelligent security and other fields. Aiming at the problem that VGG19 is easy to overfit due to the large number of fully connected layer parameters in the process of data training，CapsNet is used to replace the fully connected layer of VGG19 to realize the cascade of VGG19 and CapsNet，so as to improve the problem of overfitting during training. At the same time，the accuracy of the cascaded model on the RAF-DB dataset is improved by 5.28%. In view of the problem that the MaxPool of VGG19 feature extraction network is easy to lose information of face feature map，the SoftPool is used to replace MaxPool，so as to retain the fine-grain features of face to the maximum extent. The experiments show that the accuracy of the two improved models is 84.21% on the RAF-DB dataset and 73.16% on the FER2013 dataset，which has a good expression recognition effect.

Key words： expression recognition；VGG19；CapsNet；SoftPool

0　引言

随着深度学习的发展，人脸表情识别作为计算机视觉领域的研究热点，已越来越受到广大学者的关注，其在教室学生行为识别、“智慧安防”、“TikTok”的人脸特效、驾驶员状态监测等领域有着重要的研究意义和商业应用价值。但该研究目前存在着真实场景下表情识别难度大，以及在困难表情数据集下识别精度不高等问题[1-2]。

文献[3]针对现有表情数据集分辨率低、图片模

收稿日期：2021－08－10.

*通信作者：王敬东

糊，从而影响模型识别精度的问题，引入了注意力机制以增强模型特征提取的能力，同时针对神经网络参数大、易过拟合的问题，利用深度可分离卷积代替原始卷积，从而减少了模型的参数量，使模型在训练时能更快收敛。但该模型在训练小样本的人脸表情数据集时容易过拟合，从而影响模型精度。文献[4]提出了一个多尺度的局部注意力网络，该模型利用特征提取器提取中层特征，利用多尺度模块融合不同感受野下的特征，引导网络聚焦局部显著特征，对非正面的人脸表情识别具有一定的鲁棒性。但该模型的复杂度较高，参数量较大。文献[5]提出了一种基于视频信息的表情识别模型，该模型网络架构基于 CNN 与 LSTM，同时将面部关键点信息也作为网络的输入，从而加强网络关键区域的关注。文献[6]分别研究了 CNN 和胶囊网络在表情识别任务中的表现，证明了胶囊网络在复杂条件下的表情识别鲁棒性更好。文献[7]说明了学习特征之间的空间层次关系对于有效识别图像具有重要意义，并利用 Gabor 滤波器对图像进行预处理，使得 CapsNet 能够更好地学习图像中的空间结构关系，从而提升模型性能。文献[8]证明了 CapsNet 在分类任务上表现良好，有较小的计算成本，能降低过拟合的风险。

当利用卷积神经网络加全连接层对表情进行分类时，由于全连接层的参数量太大，增加了模型收敛的难度，使得网络在训练样本数量较少的表情数据集时容易发生过拟合，从而降低表情识别的性能。同时模型在进行前向传播时，最大池化的存在可能会丢失人脸关键区域的特征，使得表情识别的难度加大。

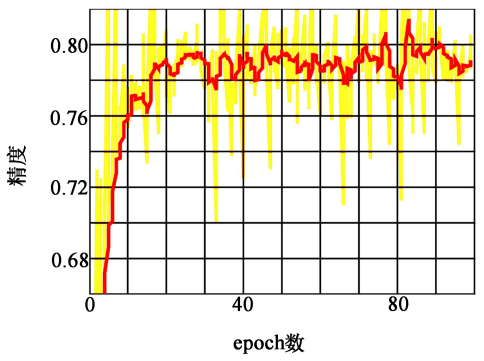
针对上述问题，本文采用 VGG19（Visual Geometry Group 19）网络来识别人脸表情，针对 VGG19 由于全连接层参数量过大而易过拟合的问题，采用 CapsNet 的胶囊层来替换全连接层，实现 VGG19 与 CapsNet 相级联；同时针对 VGG19 特征提取网络中由于最大池化层的存在而易丢失人脸关键区域信息的问题，利用 SoftPool 替换 MaxPool，最后利用实验对本文提出的改进方法加以验证。

1 VGG19 与 CapsNet 相级联的表情识别模型

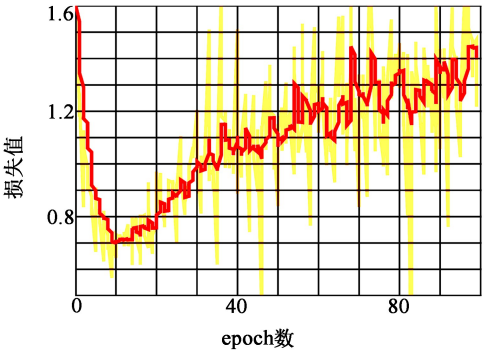
1.1 VGG19 与 CapsNet 相级联的主干网络设计

VGG 网络的全称是 Visual Geometry Group，是由牛津大学发布、主要用于图像分类领域的卷积神经网络。VGG19 结构简单，具有较深的网络结构，能提取到人脸足够的语义信息，且其以 3×3 的小卷积核为主，小卷积核使得网络能提取到更多的人脸局部细节信息，从而更有利于表情识别，因此选取 VGG19 来作为表情识别的模型。

但 VGG19 模型的分类层是三层全连接层，大大增加了模型的参数量，使得模型在训练样本数量较少的表情数据集时会出现过拟合，降低模型的泛化能力，从而降低表情识别的精度。用 VGG19 网络在 RAF-DB 上训练 100 个 epoch，每训练完一个 epoch 后利用验证集的图片来测试模型的精度和损失值。图 1 所示分别为精度和损失值随 epoch 的变化曲线。可以看出，VGG19 网络中全连接层的参数量过大，导致 VGG19 在训练 RAF-DB 数据集时出现过拟合，降低了模型的泛化能力。模型在训练集上取得了十分满意的效果，而在验证集上的精度却小于 80%，且在 epoch 大于 10 之后，模型的损失值开始上升，说明模型的泛化能力较弱，对于验证集中某些困难样本未能进行正确的识别，且这些困难样本的损失值占据了主导地位，从而使得验证集的损失值增大，说明模型的学习效果不佳。



（a） 精度变化



（b） 损失值变化

图 1 VGG19 在 RAF-DB 验证集上的精度和损失值随 epoch 的变化曲线

而胶囊网络参数相对较少，只需要用较少的数据集训练就能获得良好的分类性能，且胶囊网络能保留图像中的空间位置信息，适合于人脸这种具有空间结构信息（面部五官相对位置）的图像[9-10]，但原胶囊网络的卷积层太浅，不能提取到足够的人脸语义信息。

针对上述问题，将 VGG19 的特征提取网络和 CapsNet 的胶囊层进行级联，以实现表情识别，最后得到的 VGG19-CapsNet 模型如图 2 所示。

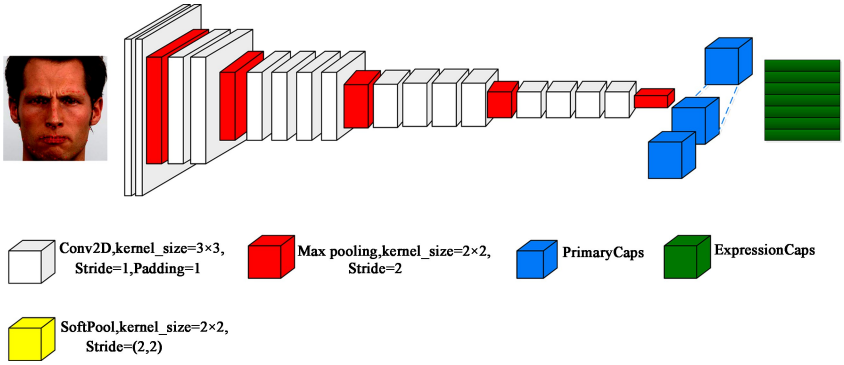


图 2　VGG19-CapsNet 模型

VGG19 与 CapsNet 的胶囊层的级联实现方式如下：网络输入的是分辨率为 224×224 的三通道彩色图片，经过卷积神经网络 VGG19 提取特征之后，在最后一个池化层输出分辨率为 7×7、通道数为 512 的特征图，作为主胶囊层 PrimaryCaps 层的输入。在主胶囊层，通过 8 次卷积核大小为 2×2×512、步长为 1、卷积核数量为 32 的卷积操作，实现将 7×7×512 大小的特征图转变为 6×6×32×8 的特征图，该特征图代表共有 6×6×32 个维度为 8 的胶囊，从而实现了将 VGG19 特征提取网络最后一个池化层输出的特征图封装为胶囊。

下一层 ExpressionCaps 层是存储高级别特征的向量，即为本文中的表情类别。本文中共有 7 类表情，所以该层为 7 个维度为 16 的胶囊。PrimaryCaps 层的 6×6×32 个维度为 8 的胶囊和 ExpressionCaps 层的 7 个维度为 16 的胶囊以全连接的形式实现向量与向量之间的连接。PrimaryCaps 层与 ExpressionCaps 层之间胶囊的运作方式示意图如图 3 所示。

假设 PrimaryCaps 层共有两个胶囊，其输出向量 $\boldsymbol{v}^1$ 和 $\boldsymbol{v}^2$，与 8×16 的矩阵 $\boldsymbol{W}^1$，$\boldsymbol{W}^2$ 相乘，得到向量 $\boldsymbol{u}^1$，$\boldsymbol{u}^2$，其中矩阵 $\boldsymbol{W}^1$，$\boldsymbol{W}^2$ 中的参数是通过反向传播学习到的。再对向量 $\boldsymbol{u}^1$，$\boldsymbol{u}^2$ 进行加权求和，得到向量 $\boldsymbol{s}$，其中权重项 c_1，c_2 也称为耦合系数，是通过动态路由算法计算得到的。最后将向量 $\boldsymbol{s}$ 送入 Squashing 激活函数，得到最后的向量 $\boldsymbol{v}$，即为 ExpressionCaps 层中的胶囊输出向量。Squashing 函数确保每个 ExpressionCaps 层的胶囊的长度在 0～1，每个胶囊的长度即为该胶囊对应的表情存在的概率。

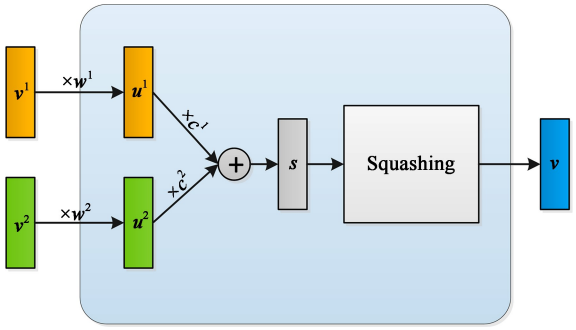
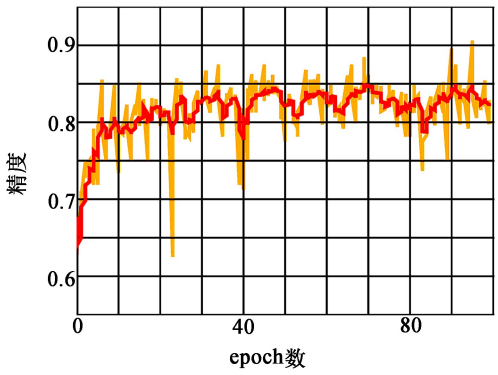
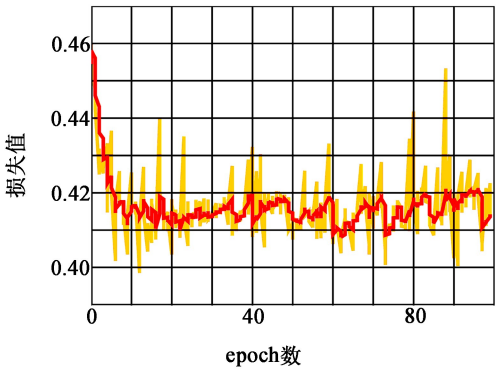


图 3　CapsNet 内部的运作方式示意图

将 VGG19 和 CapsNet 相级联的模型 VGG19-Caps 在 RAF-DB 上训练 100 个 epoch，同样每训练完 1 个 epoch 后用验证集的图片进行测试，验证过程中的精度、损失值随 epoch 的变化曲线如图 4 所示。



（a）　精度变化



(b) 损失值变化

图 4 VGG19-Caps 在 RAF-DB 验证集上的精度和损失值随 epoch 变化曲线

与原 VGG19 模型相比，VGG19 和 CapsNet 相级联的 VGG19-Caps 模型在验证过程中损失值随 epoch 变化总体呈下降趋势，说明相比于原 VGG19 模型，级联后的 VGG19-Caps 模型对于验证集有更好的泛化能力，模型有较好的学习效果，有效减小了过拟合，且验证精度相比于级联之前的模型也有了一定提高。

1.2 利用 SoftPool 对池化方式进行改进

CNN 中的池化层可以对特征图起到降采样的作用，同时减少模型的参数，降低模型过拟合的风险[11-12]。VGG19 中的最大池化可以保留池化内核中的局部最大值，凸出内核中的显著特征。但利用最大池化进行特征映射时可能会丢失人脸的关键区域信息，从而降低模型对表情的识别精度。针对这一问题，本文利用 SoftPool[13] 替换 VGG19 中的最大池化层，SoftPool 对感受野内的所有像素以加权求和的方式映射到网络的下一层中，其中用到的是 softmax 的加权方法，该方法相对于最大池化而言，在对特征图进行降采样的同时，还能保留特征图中更多的人脸细粒度信息，适用于表情识别的特征提取网络。SoftPool 的实现方式如图 5 所示。

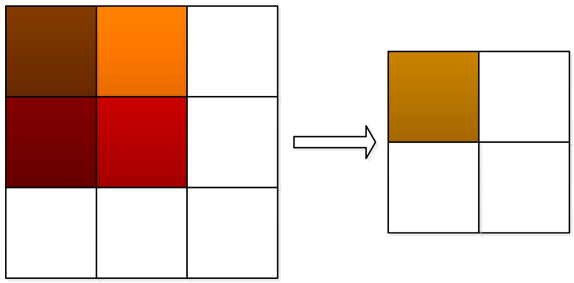


图 5 SoftPool 的实现方式

设左图 SoftPool 所对应的感受野中每个像素的激活值为 a_i，池化内核定义为 $\mathcal{R}$，感受野内每个像素的权重为

$$w_i = \frac{e^{a_i}}{\sum_{j \in \mathcal{R}} e^{a_j}} \tag{1}$$

则右图经过 SoftPool 后的输出值为

$$\tilde{a} = \sum_{i \in \mathcal{R}} w_i \times a_i \tag{2}$$

用 SoftPool 替换最大池化后的模型 VGG19-Caps-Soft 如图 6 所示。

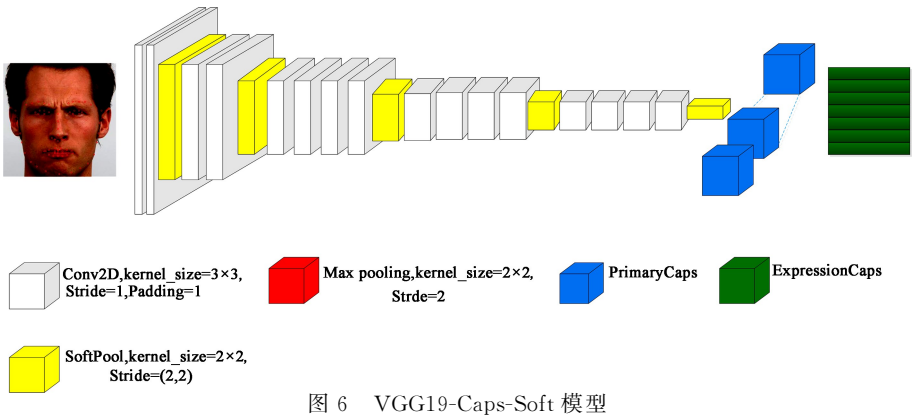


图 6 VGG19-Caps-Soft 模型

为了直观地感受 SoftPool 保留人脸细粒度特征的性质，将人脸图像先后送入未添加 SoftPool、添加 SoftPool 的模型，并将特征提取网络 VGG19 的第三个池化层输出的特征图按通道取平均之后，进行可视化，得到的可视化结果如图 7 所示。

图 7 中从左到右依次为原图、未添加 SoftPool 的模型输出的特征图、添加 SoftPool 的模型输出的特征图。可以看出，相比于使用最大池化的模型，使用 SoftPool 的模型能最大程度上保留人脸的细节信息以及细粒度特征。

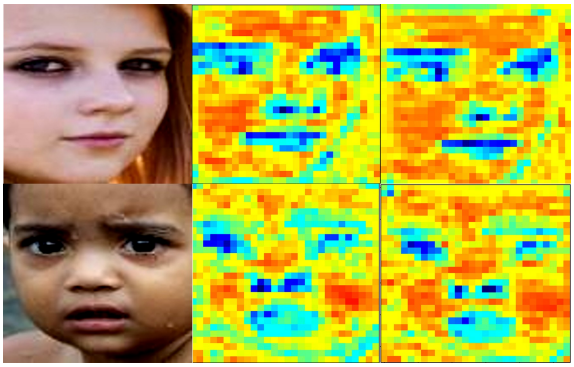


图 7　特征图可视化

2　实验结果与分析

本文实验所采用的平台为台式服务器，配置如下：Ubuntu16.04 的系统，Intel 的 CPU 和 NVIDIA GTX 1080Ti 的 GPU。本文算法在 Pytorch 深度学习框架下使用 Python 语言编程实现，训练参数设置：迭代批次：*epoch* = 100，每批次的大小：batch size 设为 14，优化器选用 AdamOptimer，初始学习率为 0.001。

2.1　表情数据集

本文实验所采用的数据集为 RAF-DB 数据集[14]和 FER2013 数据集。RAF-DB 数据集是从互联网上搜集到的大规模的人脸表情数据集，其包含了超过 3 万张人脸表情图像，该数据集中的人分布于各个年龄段，且不同国籍，部分图片存在一定程度的头部偏转。本文选取 RAF-DB 数据集中的基础类型图片来进行研究，其中训练集图片 12 271 张，测试集图片 3 068 张，包含 7 类表情标签：surprise，fear，disgust，happy，sad，angry 和 neutral，每张表情图片对应一种标签类型。RAF-DB 数据集展示如图 8 所示。



图 8　RAF-DB 数据集

FER2013 数据集由分辨率大小为 48×48 的灰度图组成，由于该表情数据集的组成较为复杂，存在一些不属于人脸表情的图像，如动漫图像、非人脸图像等。本文对该数据集中的错误图像进行人为剔除，剔除后的数据集图片总量为 26 797 张，其中训练集 23 925 张，测试集 2 872 张。FER2013 数据集展示如图 9 所示。



图 9　FER2013 数据集

2.2　实验结果分析

为了验证本文所作的改进，本文选取了以下改进前后的三种网络进行消融实验：未做任何变动的 VGG19、将 VGG19 的全连接层替换为胶囊层的 VGG19-Caps、将 VGG19-Caps 特征提取网络里的 MaxPool 替换为 SoftPool 的 VGG19-Caps-Soft 模型。在不加预训练权重的情况下，让上述三个模型分别在 RAF-DB 训练集上训练 100 个 epoch，最后将训练得到的模型在测试集上进行测试，得到的消融实验结果如表 1 所示。

表 1　模型在 RAF-DB 数据集上的消融实验

模型	CapsLayer	SoftPool	准确率/%
VGG19			78.57
VGG19-Caps	√		83.85
VGG19-Caps-Soft	√	√	84.21

上表中打勾项表示模型中存在该模块，未打勾则说明模型中不存在该模块。可以看出，将 VGG19 的全连接层替换为胶囊层 CapsLayer 实现分类后，模型提升的精度最高，为 5.28%。通过上述消融实验可以看出，本文所作出的上述两种改进对于表情识别精度的提升都起到了一定作用，且两种改进加在一起时效果最好，使得表情识别在 RAF-DB 数据集上的精度达到了 84.21%。

为了对比原始 VGG19 模型和本文改进后

VGG19-Caps-Soft 模型对于不同表情的识别情况，利用在 RAF-DB 训练集上训练的上述两个模型，在 RAF-DB 测试集上进行测试，得到的测试结果由图 10～11 两个混淆矩阵表示。

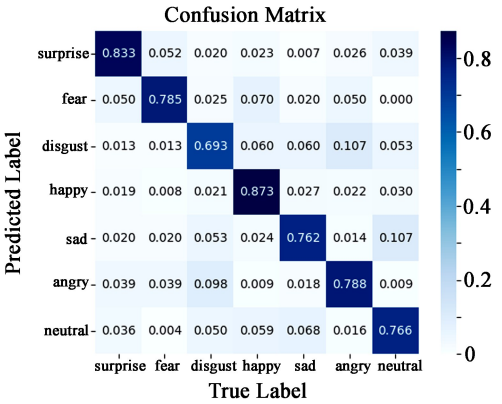


图 10 VGG19 在 RAF-DB 测试集上得到的混淆矩阵

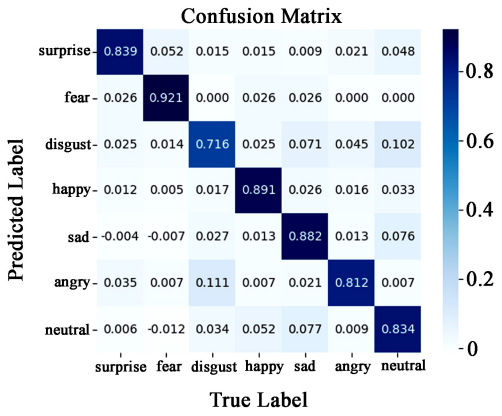


图 11 VGG19-Caps-Soft 在 RAF-DB 测试集上得到的混淆矩阵

混淆矩阵的行表示预测结果，列表示真实标签。对角线上的元素即为预测正确的结果，对角线以外的元素即为预测错误的结果。为了清晰地描述模型预测不同表情的准确率（预测正确的个数占预测总个数的比值），将混淆矩阵归一化显示。

由图 10～11 可以看出，改进后的模型有效提高了 fear 和 sad 这两种表情的识别精度，分别提高了 13.60%和 12.00%，说明了模型的改进有效提升了这两类表情的识别精度。对于 neutral 表情，精度提升了 6.8%。对于其他类别的表情，精度提升并不明显，这也是本课题后续研究的方向。改进前后所有表情类别的平均精度提高了 5.64%，证明了本文改进方法的有效性。

将本文训练好的最优模型和几个典型表情识别算法在 RAF-DB 数据集和 FER2013 数据集上进行测试比较，得到的准确率如表 2～3 所示。从表中可以看出，与典型算法相比，本文改进的 VGG19-Caps-Soft 模型在 RAF-DB 数据集和 FER2013 数据集上都具有较高的精度，准确率分别达到了 84.21%和 73.16%，说明改进取得了不错的效果。

表 2 不同表情识别算法在 RAF-DB 数据集上的比较

算法	准确率/%
HOG	61.46
Gabor	65.17
Occams Razor[15]	80.00
pACNN[16]	83.05
VGG19-Caps-Soft	84.21

表 3 不同表情识别算法在 FER2013 数据集上的比较

算法	准确率/%
DenseNet	65.01
MobileNet V2	67.60
Zhou[17]	70.91
Tian[18]	72.64
VGG19-Caps-Soft	73.16

3 结论

本文针对 VGG19 的全连接层参数量太大而使得模型训练时容易过拟合的问题，用参数更少、分类性能良好、且更适合于人脸图像的胶囊网络的胶囊层对全连接层进行替换，构成 VGG19 和 CapsNet 相级联的模型，实验证明改进后模型的过拟合问题得到了明显改善，且表情识别精度有明显提高。针对 VGG19 特征提取网络中的最大池化层会丢失大量人脸特征图信息的问题，本文用 SoftPool 对最大池化层进行替换，该方法能够保证在对特征图进行降采样的同时一定程度上减少特征图中信息的损耗。最后通过消融实验证明了两种改进方法都起到了一定效果，且两种改进方法叠加在一起时识别精度最高。

参考文献：

[1] 陈欣键，姜 飞，申瑞民，等. 基于区域与分类回归融合的 AU 强度估计算法[J]. 计算机工程，2019，45(1)：186-191.
Chen Xinjian，Jiang Fei，Shen Ruimin，et al. Au intensity estimation algorithm based on region and classification regression fusion[J]. Computer Engineering，2019，45(1)：186-191.

[2] 冯 杨，刘 蓉，鲁 甜. 基于小尺度核卷积的人脸表情识别[J]. 计算机工程，2021，47(4)：262-267.
Feng Yang，Liu Rong，Lu Tian. Facial expression recognition based on small-scale kernel convolution [J]. Computer

Engineering，2021，47(4)：262-267.

[3] Gao H，Ma B. A robust improved network for facial expression recognition[J]. Frontiers in Signal Processing，2020，4(4)：4.

[4] Zhao Z，Liu Q，Wang S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild[J]. IEEE Trans. on Image Processing，2021，30：4973-4984.

[5] Hasani B，Mahoor M H. Facial expression recognition using enhanced deep 3D convolutional neural networks[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops，2017：30-40.

[6] Li D，Zhao X，Yuan G，et al. Robustness comparison between the capsule network and the convolutional network for facial expression recognition[J]. Appl. Intelligence，2021，51(4)：2269-2278.

[7] Ayidzoe M A，Yu Y，Mensah P K，et al. Gabor capsule network with preprocessing blocks for the recognition of complex images[J]. Machine Vision and Applications，2021，32(4)：1-16.

[8] Mensah P K，Weyori B A，Ayidzoe M A. Evaluating shallow capsule networks on complex images[J]. Inter. J. of Information Technol.，2021，13(3)：1047-1057.

[9] Li Wei，Abtahi F，Zhu Z，et al. EAC-Net：Deep nets with enhancing and cropping for facial action unit detection[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence，2018，40(11)：2583-2596.

[10] Cao Shan，Yao Yuqian，An Gaoyun. E2-capsule neural networks for facial expression recognition using AU-aware attention[J]. IET Image Processing，2020，14(11)：2417-2424.

[11] 邓楚婕. 基于卷积神经网络的人脸表情识别方法研究[D]. 广州：华南理工大学，2020.
Deng Chujie. Research on facial expression recognition based on CNN[D]. Guangzhou：South China University of Technology，2020.

[12] 王莹莹. 基于卷积神经网络的人脸表情识别问题的研究[D]. 济南：山东大学，2020.
Wang Yingying. Research on facial expression recognition based on convolutional neutral network[D]. Ji'nan：Shandong University，2020.

[13] Stergiou A，Poppe R，Kalliatakis G. Refining activation downsampling with Softpool[J]. arXiv preprint：2101.00440.

[14] Li S，Deng W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition[J]. IEEE Trans. on Image Processing，2018，28(1)：356-370.

[15] Vielzeuf V，Kervadec C，Pateux S，et al. An Occam's Razor view on learning audiovisual emotion recognition with small training sets[C]// Proc. of the 20th ACM Inter. Conf. on Multimodal Interaction，2018：589-593.

[16] Li Y，Zeng J，Shan S，et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Trans. on Image Processing，2018，28(5)：2439-2450.

[17] Zhou Jiancan，Jia Xi，Shen Linlin，et al. Improved softmax loss for deep learning-based face and expression recognition[J]. Cognitive Computation and Systems，2019，1(4)：97-102.

[18] Tian Yi，Wen Zhiwei，Xie Weicheng，et al. Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition[C]// 2019 IEEE Inter. Conf. on Image Processing (ICIP)，2019：46-50.

作者简介：

杨　双(1996—)，男，硕士研究生，主要研究方向为数字图像处理；

王敬东(1966—)，男，副教授，硕士生导师，主要研究方向为数字图像处理。