计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):259−266 [doi: 10.15888/j.cnki.csa.008273]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于图卷积多标签学习的复合人脸表情识别①

武中华

(江苏大学 计算机科学与通信工程学院, 镇江 212013)
通信作者: 武中华, E-mail: 2200680719@qq.com

摘　要: 传统的人脸表情识别方法主要针对六类基本人脸表情, 但在现实场景下, 存在更加丰富的由基本人脸表情组合而成的复合人脸表情, 原先识别基本人脸表情的工作难以去识别复合人脸表情, 并且复合人脸表情的数据集缺乏足够的训练数据. 针对该问题, 提出基于图卷积多标签学习的复合人脸表情识别方法. 通过特征提取网络提取到人脸表情的全局特征和感兴趣区域的局部特征, 使用基本和复合人脸表情之间的先验知识和数据驱动方式, 构建出表情类别关系图, 利用图卷积网络来学习到表情类别分类器, 最后进行复合人脸表情识别. 在 RAF-DB 和 EmotioNet 数据集上的实验结果表明, 与 VGG19 和 ResNet50 等方法相比, 该方法可以使得复合人脸表情识别率取得约 4%~5% 的提升.

关键词: 复合人脸表情识别; 图卷积网络; 知识图; 词向量; 多标签学习

引用格式: 武中华.基于图卷积多标签学习的复合人脸表情识别.计算机系统应用,2022,31(1):259–266. http://www.c-s-a.org.cn/1003-3254/8273.html

GCN in Multi-label Learning for Compound Facial Expression Recognition

WU Zhong-Hua

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Traditional facial expression recognition (FER) methods have focused on the six basic facial expressions. However, compound facial expressions are also used by humans in the real world. Compound facial expression means that it is a combination of the basic facial expressions. However, the traditional methods of recognizing basic facial expressions are unable to handle compound facial expressions. Moreover, the compound facial expression datasets have insufficient training data. To address the difficulties in the compound FER, this study proposed a graph convolutional network in multi-label learning for compound facial expression recognition (GCN-ML-CFER). The global features of facial expression and the local features of the regions of interest were extracted by the feature extraction network. According to the prior knowledge of basic and compound facial expressions, a relationship graph of facial expression categories was constructed by a data-driven method. The expression category classifiers learn the graph via a graph convolutional network (GCN). Finally, compound FER was carried out by the classifiers. Experiments were conducted on the RAF-DB and EmotioNet datasets. The results show that this method achieves a 4%–5% increase in the compound FER accuracy compared with those of the VGG19 and ResNet50 methods.

Key words: compound facial expression recognition (FER); graph convolutional network (GCN); knowledge graph (KG); word embedding; multi-label learning

① 收稿时间: 2021-03-30; 修改时间: 2021-04-29; 采用时间: 2021-05-07; csa 在线出版时间: 2021-12-17

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在现实生活中, 人脸表情是仅次于语气之后必不可少的情感交流手段[1]. 人脸表情识别能让计算机有效表达人类的情感信息, 是人工智能领域中的重要组成部分. 人脸表情识别是将人脸表情图像识别为不同的表情类型, 如愤怒、高兴、悲伤、惊讶、厌恶和恐惧等等[2]. 近年来, 随着人工智能研究领域的不断发展, 人脸表情识别也因其重要性而受到广泛关注.

目前, 人脸表情识别方法划分为 3 个主要步骤, 分别是预处理, 人脸表情特征提取和人脸表情分类. 在人脸表情特征提取中, 根据特征提取方式不同分为手工特征和学习型特征, 前者是通过手工设计的算法进行提取, 后者是通过深度学习模型进行提取. 对于手工特征, 可以进一步分为基于纹理的特征, 如局部二值模式 (local binary pattern, LBP)、Gabor 小波变换; 基于几何的特征, 如尺度不变特征变换 (scale-invariant feature transform, SIFT) 和基于多种手工特征得到的混合特征. 而大多数学习型特征都是基于神经网络自动进行学习[3,4], 如卷积神经网络 (convolutional neural network, CNN), 深度神经网络 (deep neural network, DNN), 循环神经网络 (recurrent neural network, RNN)和生成对抗网络 (generative adversarial network, GAN). 人脸表情分类的方法则有支持向量机 (support vector machine, SVM)、隐马尔科夫模型 (hidden Markov model, HMM)、K 最近邻算法 (K-nearest neighbor, KNN) 和混合分类器模型等[5,6].

传统的人脸表情识别仅限于识别 6 种基本人脸表情, 即愤怒、高兴、悲伤、惊讶、厌恶和恐惧. 然而, 现实生活中人类情感变化非常复杂, 表现出来的人脸表情类别大大高于早期定义的 6 种基本表情[7]. 复合人脸表情的提出为人脸表情识别开辟了一个新的领域, 可以将计算机视觉和人工智能的研究提高到一个新的高度. 复合人脸表情通常是来自于没有任何控制条件下的真实场景, 而大部分公开的自然环境下的人脸表情数据集只包含基本表情, 而少数包含复合人脸表情的数据集也缺乏足够的训练数据. 近年来, 有一些基于复合人脸表情识别的研究, Benitez-Quiroz 等人[8] 提出了利用检测表情中的面部运动单元 (action unit, AU) 来识别复合人脸表情, 然而此方法需要显著提升表情中 AU 的检测性能才能有效识别复合人脸表情. Li 等人[9] 改进了基础的深度卷积神经网络 (deep convolutional neural network, DCNN) 提出一种新的模型 (deep locality-preserving CNN, DLP-CNN) 来进行复合人脸表情识别, 该方法大大增强了识别能力. 但是, 复合人脸表情数据集训练样本的不足, 人工标注费时费力, 因此, 目前的研究还是主要集中在基本人脸表情的识别.

单标签学习中, 每个样本只属于一个标签且标签之间两两互斥, 而在多标签学习中, 一个样本可以对应多个标签, 且各个标签之间通常具有一定的联系[10]. 在现实生活中, 数据的复杂性导致单标签学习已经无法满足研究方法的要求, 因为真实的对象往往具有多义性, 所以多标签学习逐渐得到了广泛的关注. 运用面部动作编码系统 (facial action coding system, FACS)[7] 对所有人脸表情中出现的人脸面部运动单元 (AU) 进行研究发现, 复合人脸表情一般是由两种基本人脸表情组合而成的, 如惊喜 (happily surprised), 其是由高兴 (happily) 和惊讶 (surprised) 两个基本表情组合而成, 所以复合人脸表情识别可以视为一个多标签分类问题.

现实世界中的诸多问题都是用图的形式来表示, 近年来, 由于图卷积网络能够解决图的卷积问题得到了巨大的发展[11,12]. Wang 等人[13] 利用图卷积进行零样本图像识别时, 考虑到可见类别和不可见类别之间的关系, 转移从可见类别中学习到的知识来描述不可见类别, 大幅度提高了零样本识别的性能. Chen 等人[14] 通过构建知识图来捕获标签间的依赖关系, 将图卷积应用在多标签图像识别上, 也取得了巨大的成功. Zhang 等人[15] 利用上下文信息来构建情感关系图, 再利用图卷积网络来学习情感关系以推理情绪状态, 获得不错的效果. Li 等人[16] 在人脸面部单元识别中, 先利用先验知识构造了 AU 关系图, 再使用 GGNN 在图上进行信息传播来得到 AU 的特征, 最后进行 AU 识别. 表明了人脸面部单元识别中使用图神经网络的有效性.

我们将复合人脸表情识别视为多标签分类问题, 通过复合人脸表情类别之间的联系来构建人脸表情类别关系知识图, 为了更好的获得表情之间的关系, 我们提出了一种基于图卷积网络多标签学习的复合人脸表情识别方 (graph convolution network in multi-label learning for compound facial expression recognition, GCN-ML-CFER), 来更好的实现对复合人脸表情的识别.

1 基于图卷积多标签学习的复合人脸表情识别

图 1 所示为整体的网络结构, 基于图卷积多标签

© 中国科学院软件研究所 http://www.c-s-a.org.cn

学习的复合人脸表情识别模型 (GCN-ML-CFER) 主要分为 3 个部分: 1) 以 VGG19 网络模型为骨架, 再利用提供的人脸面部关键点来对感兴趣区域 (region of interest, ROI)[16,17] 进行学习, 最后提取人脸表情的特征. 2) 通过面部动作编码系统 (FACS) 对所有人脸表情中出现的人脸面部运动单元 (AU) 进行分析, 得到人脸表情类别之间的关系. 再通过数据驱动的方式, 挖掘人脸表情类别的标签在数据集中的共现模式, 使用条件概率的形式对标签的依赖性关系进行建模, 得到人脸表情类别关系图, 图卷积网络作用在关系图上进行分类器学习. 3) 通过提取的人脸表情特征与学习到的分类器进行复合人脸表情预测.

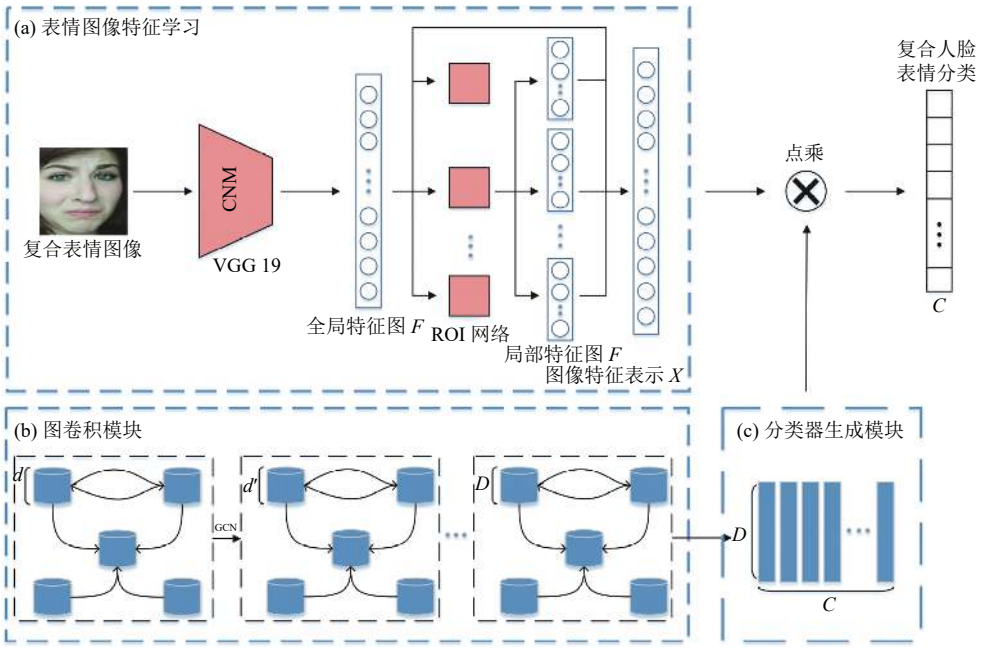


图 1 GCN-ML-CFER 模型框架

1.1 人脸表情特征提取模块

面部动作编码系统 (FACS) 根据人脸解剖学的特点用人脸面部运动单元 (AU) 的变化来描绘不同的表情这种描述方式几乎可以表现所有的面部表情, 是目前标准的表情划分参照体系. 人脸表情的发生是基于人脸面部运动单元 (AU) 的变化, 所以为了获得更加显著的人脸表情特征, 我们利用提供的人脸面部关键点来对感兴趣区域进行特征提取, 这种全局和局部特征的结合, 可以很好的表示人脸的表情.

我们选择 VGG19 作为我们的骨架网络, 如图 1(a) 所示, 通过 VGG19 我们可以得到人脸面部表情的全局特征图 F. 接着我们根据提供的 5 个人脸面部关键点置, 缩放映射到全局特征图的关键点位置, 以此位置为中心, 划分出感兴趣区域, 使用 ROI 网络来对这 5 个感兴趣区域进行特征提取, 从而得到局部的人脸表情特征.

$$f_i = ROI(F, l_i), i \in \{1,2,\cdots,5\} \quad (1)$$

其中, l为提供的人脸面部关键点, ROI 为区域特征提取网络, f为从感兴趣区域提取到的特征. 将连接起来的全局和局部特征作为我们从人脸表情图像中提取得到的表情特征.

$$X = g(F, f_i), i \in \{1,2,\cdots,5\} \quad (2)$$

其中, g为特征连接, X 为最后的维度为 600 的人脸表情特征.

1.2 表情类别知识图构造

考虑到复合表情之间具有一定的相关性, 捕获和利用这些相关性可以提升复合表情的分类性能. 拓扑结构的图拥有对于复杂系统的强表现力, 同时具有很强的推理能力, 因此将人脸表情类别之间的关系构造成图的形式, 可以很好的进行复合人脸表情识别.

我们用 V 来表示图中节点的集合, 具体来说, 每种基本表情类别分别对应图的一个节点, 即$v \in V$, 图中的每个节点表示为标签的词嵌入. 词嵌入是一种将文本中的词转表示数字向量的方法, 向量中的每一个维度可视为对应特定的语义信息, 在词嵌入空间中, 语义相

© 中国科学院软件研究所 http://www.c-s-a.org.cn

关和相近的概念词向量也彼此接近.

图中节点间的关系我们用边 E 来表示, 如惊喜 (happily surprised) 这个复合表情, 它在图中体现为高兴 (happily) 代表的节点和惊讶 (surprised) 代表的节点通过边来进行连接. 根据面部动作编码系统 (FACS) 和复合表情的标签, 我们初步可以得到哪些基本表情之间是有关系的, 也就是图中哪些节点是通过边相连的. 同时, 我们再通过数据驱动的方式来进一步表示图中节点间关系的强度, 即通过挖掘数据集中不同复合表情的数量, 来对图中相连节点之间关系进行调整.

我们以条件概率的形式对节点间关系的强度进行建模. 即$P(L_j|L_i)$, 它表示的是出现标签L_i时出现标签L_j的概率, 需要注意的是, $P(L_j|L_i)$不等于$P(L_i|L_j)$.

复合表情可以视为基本表情的标签对. 首先我们对训练集中的所有复合表情进行计数, 得到矩阵$M \in R^{C\times C}$, 其中, C为基本表情的个数, M_{ij}表示基本表情标签L_i和L_j一同出现的次数, 也就是, 这两个基本表情组成的复合表情出现的次数.

再利用$P_{ij} = M_{ij}/N_i$得到条件概率矩阵$P \in R^{C\times C}$, 其中N_i表示基本表情标签L_i在数据集中出现的次数, $P_{ij} = P(L_j|L_i)$.

在图卷积后, 节点的特征为节点自身特征与相邻节点特征的加权和, 对于图卷积可能导致的过渡平滑问题, 即节点特征可能变得相似, 以至于不同类别的节点可能变的难以区分, 为了缓解这个问题, 我们对条件概率矩阵进行一定的改进, 首先对于可能出现的噪声边通过阈值t来进行限制.

$$A_{ij} = \begin{cases} 0, & \text{if } P_{ij} < t \\ 1, & \text{if } P_{ij} \geq t \end{cases} \tag{3}$$

接着, 在更新节点特征时, 有一个固定的权重对节点本身的特征, 而相邻节点的特征由其分布决定, 最后邻接矩阵 A 表示为:

$$A = \begin{cases} p \Big/ \sum_{j=1,i\neq j}^{C} A_{ij}, & \text{if } i \neq j \\ 1-p, & \text{if } i = j \end{cases} \tag{4}$$

其中, A 是邻接矩阵, 而p是分配给节点本身和其相邻节点的权重, 当p趋近于 1 时, 节点本身的特征将不会着重考虑, 主要使用其相邻节点的特征. 当p趋近于 0 时, 其相邻节点的特征将不会着重考虑, 主要考虑节点本身的特征.

1.3 图卷积模块的分类器学习

图卷积在学习过程中能够融合图结构信息, 可以将来自相邻节点的有效信息集成到节点自身当中, 因此, 我们使用图卷积从表情类别知识图中学习表情类别分类器, 如图 1(b) 和图 1(c) 所示. 给定的图是一个具有C个节点且每个节点的特征维度为d, 从而得到图的特征矩阵$H^0 \in R^{C\times d}$. 其中节点的初始特征为相对应表情标签的词向量表示. 表情类别知识图用邻接矩阵$A \in R^{C\times C}$表示. 我们采用简单的传播规则进行图卷积.

$$H^{l+1} = \sigma(\hat{A}H^lW^l) \tag{5}$$

其中, σ为 ReLU 激活函数, $\hat{A}$是对邻接矩阵 A 进行归一化后的矩阵, H^l是第l层的节点特征表示, 首层的节点特征表示为$H^{(0)}$, 通过图卷积将图的节点特征矩阵更新为$H^{l+1} \in R^{C\times d'}$, 可以通过多层的图卷积来学习和建模节点间复杂的关系, W^l是第l层待学习的权重参数, 最后通过图卷积后的输出为$Z \in R^{C\times D}$, D 与人脸表情特征 X 的维度相同.

经过图卷积模块得到的$Z \in R^{C\times D}$就是我们学习到的分类器, 将其应用到人脸表情特征上, 就可以得到表情类别预测的分数:

$$\hat{y} = ZX \tag{6}$$

人脸表情图像的标签为$y \in R^C$, 其中$y_i = \{0,1\}$表示人脸表情类别标签i是否出现在图像中. 整个网络用传统的多标签分类损失进行训练:

$$L = \sum_{c=1}^{C} y_c \log(\varphi(\hat{y}_c)) + (1-y_c)\log(1-\varphi(\hat{y}_c)) \tag{7}$$

其中, φ是Sigmoid函数.

2 基于图卷积多标签学习的复合人脸表情识别方法

本文在 2 个数据集上进行复合人脸表情识别实验.

RAF-DB[9] 是目前最大公开可用的真实情感人脸数据集, 它拥有 15 339 张 7 种基本表情图像和 3 954 张 11 种复合人脸表情图像. 本文使用 11 种复合人脸表情, 采用数据集提供的 3 162 张训练集图像和 792 张测试图像.

EmotioNet[8] 是自然环境下大型人脸表情数据集, 它拥有 2 478 张带有人脸表情标签的图像, 由于我们工作集中在复合人脸表情识别上, 同时选择有明确基本表情组成的复合人脸表情类别, 最后我们从中获取了 1 220 张复合人脸表情图像, 其中训练集图像为 980 个, 测试集图像为 240 个.

在实验设置方面, 首先, 我们采用 4 层图卷积网络, 每层维度为 350, 400, 500, 600. 表情类别知识图构造

© 中国科学院软件研究所 http://www.c-s-a.org.cn

中, 我们选择的是 300 维度的 GloVe[18] 词向量作为每个节点的初始化, 图中边的构造中, 我们的参数设置为 p=0.3, t=0.2. 在人脸表情特征提取模块, 我们采用 LeakyReLU=0.2 激活函数, 预训练的 VGG19 为主干网络, 在训练过程中, 输入人脸表情图像大小归一化到为 100×100, 最后得到的图像特征维度为 600, 与最后图卷积后的节点维度一致. 采用 SGD 优化算法, momentum 为 0.9, 学习率初始化设置为 0.01, 每 30 个 epoch 学习率衰减 10 倍. 整个网络构建使用的是 Python 3.6, CUDA10.2, PyTorch 1.3.1.

2.1 特征提取模型选择

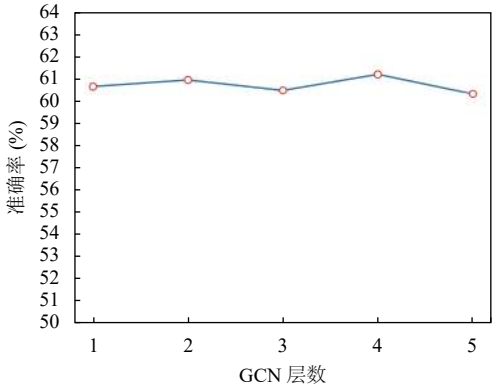
为了选择合适的特征提取模型, 在 RAF-DB 这个数据集上, 对几个目前流行的深度学习模型的识别准确率进行了对比, 即 baseDCNN[9], ResNet18, ResNet34, ResNet50, ResNet101[19] 和 VGG19[20]. 其中, baseDCNN 是 RAF-DB 数据库中基准方法 DLP-CNN 的特征提取模型, DLP-CNN 能够提高对学习到的特征的识别能力, 可以比拟于其它最优的方法. 所有的模型都是用 RAF-DB 的训练集数据进行训练, 在测试集上进行测试, 结果如表 1 所示, 我们使用的模型除了 baseDCNN 外, 其它都是经过 ImageNet[21] 预训练过后的模型. 从表中可以看出, 其它模型的识别率相对 VGG19 来说, VGG19 的结果最好, 因此, 后续的试验以 VGG19 作为选择的特征提取模型.

表 1 不同模型的识别准确率比较 (%)

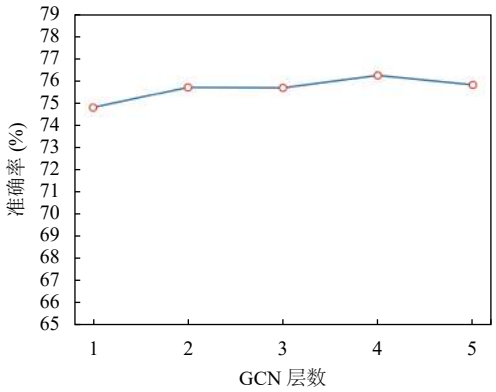
对比方法	RAF-DB	EmotioNet
baseDCNN	46.717	60.581
ResNet18	56.818	74.167
ResNet34	55.808	73.333
ResNet50	55.454	73.750
ResNet101	54.672	74.583
VGG19	57.071	75.417

2.2 卷积层数的选择

我们展现了不同卷积层数对模型识别率的影响, 对于 3 层图卷积网络, 输出的维度分别是, 400, 500, 600, 对于 4 层图卷积, 输出维度为 350, 450, 550, 600, 对于 5 层图卷积, 输出维度为 350, 400, 450, 500, 600. 通过图 2 中的结果展示, 随着图卷积数目的增加, 复合表情识别率先上升后下降, 在使用 4 层图卷积的情形下识别率最高. 可能的原因是, 在使用更多的图卷积层时, 节点之间的多次传播导致了过平滑, 使得节点间的区分性降低, 导致识别率的降低, 而我们为了缓解过平滑, 在知识图的构造过程中, 设置了t来限制节点之间边的连接, 设置p来分配给节点本身和其相邻节点的权重, 一定程度缓解了使用更多的图卷积层而出现的过平滑, 所以才会出现随着图卷积层数的变化, 复合表情识别率也出现了先上升后下降的变化, 而且变化幅度不大.



(a) RAF-DB 数据集



(b) EmotioNet 数据集

图 2 两个数据集下不同 GCN 层数的准确率

2.3 不同词向量选择的影响

在表情类别知识图构造中, 运用词向量来对图中节点进行初始化, 我们调查了几个不同的词向量表示, 包括 GloVe, GoogleNew[22] 和 FastText[23]3 个词向量表示. 图 3 展示了这 3 种词向量对实验结果的影响, 对比于其它的词向量, GloVe 词向量下模型的识别率相对较高. 我们发现, 3 种不同的词向量下实验结果差别不是很大, 表明我们模型的识别率受词向量的影响较小. 同时运用更加合理准确的词向量能够得到更好的结果, 原因可能是从丰富语料中学习到的词向量包含了丰富的语义信息, 我们的模型能够利用这种有效的语义信息来提升对复合人脸表情识别的准确率.

2.4 t取值分析

在表情类别知识图构造中, 邻接矩阵中的t是一个阈值, 来决定图中两个节点是否进行连接. $t \in \{0,0.1,$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

0.2,⋯,0.9,1}, 其结果如图 4 所示. 我们发现, 当t取值为 0 时, 表示所有的节点进行连接, 随着t值的增加, 减少了一些干扰的边, 使得识别的准确率不断的增加, 然而, 当太多的边删减之后, 节点之间的关系不能很好的学习到, 导致准确在不断的下降. 我们从图中发现在 RAF-DB 数据集中, $t=0.2$时, 复合表情的识别率最好, 而在 EmotioNet 数据集中, $t=0.4$时, 复合表情的识别率最好. 出现此类情况的原因可能是, 不同的数据集所拥有的的复合表情的数目不同, 而根据数据驱动而构造的知识图也因此受到影响, 导致不同数据集下合适的t值是不同的.

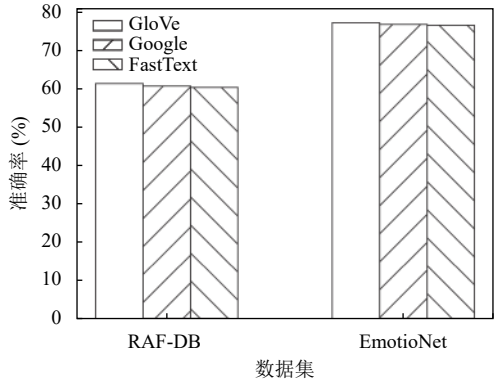


图 3 两个数据集下不同词向量的准确率

2.5 p取值分析

在表情类别知识图构造中, 邻接矩阵中的p是分配给节点本身和其相邻节点的权重. 为了发现不同p值构造的知识图对复合表情识别的影响, 我们应用$p\in\{0,0.1,0.2,\cdots,0.9,1\}$, 结果如图 5 所示, 我们能发现当$p=0.3$时, 它能取得最好的结果. 如果$p$值太小, 图中节点不能从邻接节点中学习到有效的信息, 如果p值太大, 它将不会保持自身的特征, 导致出现过平滑现象.

2.6 ROI 网络的影响

我们根据 RAF-DB 数据集提供的 5 个人脸面部关键点位置, 划分出感兴趣区域, 使用 ROI 网络来对这 5 个感兴趣区域进行特征提取, 将得到的局部的人脸表情特征和全局人脸特征进行结合得到最后的人脸表情特征. 为了验证 ROI 网络的有效性, 我们从模型中移除 ROI 网络, 直接进行复合人脸表情识别, 我们将缺失了 ROI 网络的模型称为 GCN-ML-CFER-ROI. 由于 EmotioNet 数据集中没有提供准确的人脸面部关键点位置, 所在 RAF-DB 数据集中进行比较. 对比结果如表 2 所示. 我们发现, 在 RAF-DB 数据集中, ROI 网络的使用提升了复合人脸表情的识别率, 提升了大约 1.3% 左右, 原因很可能是通过 ROI 网络, 我们提取到了更有效的人脸表情特征, 从而使得整个模型的复合人脸表情准确率得到了提升.

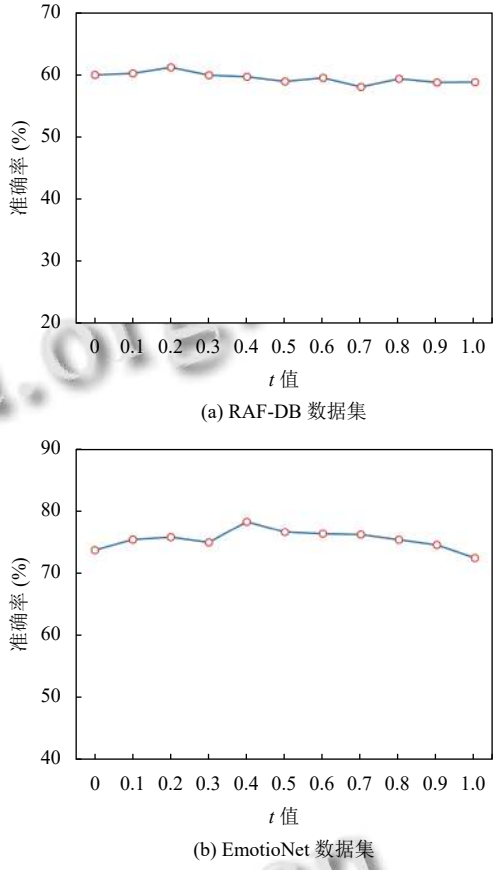


图 4 两个数据集下不同 t 值的准确率

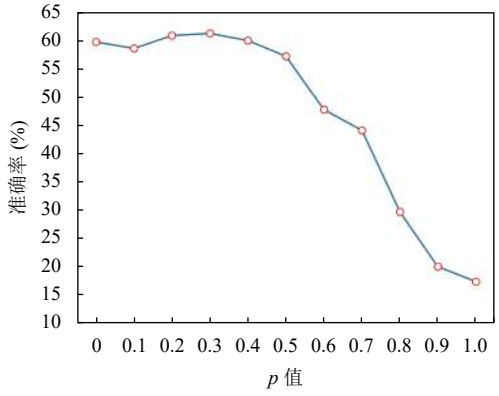
2.7 与其它方法比较

实验与目前的主流研究方法作对比. 表 1 给出了对比方法的准确率结果. 表 1 中的对比方法是在单标签学习基础上进行的, 复合人脸表情图像对应复合人脸表情类别, 即一张人脸表情图像对应一个标签. 表 3 中是我们提出的基于图卷积多标签的复合人脸表情识别模型 GCN-ML-CFER 的准确率结果. 从表 3 中可以明显看出多标签学习下对复合人脸表情识别的准确率明显高于单标签学习下的准确率.

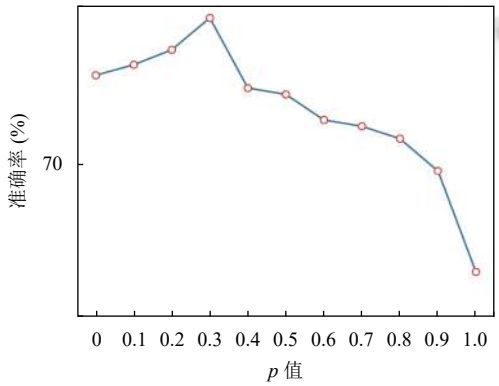
表 3 给出了在多标签学习中, 我们模型在不同主干网络下的准确率结果, 从中可以看出: 将我们模型的提取人脸表情特征的主干网络替换, 整个模型的复合人脸表情识别率都高于对应的原先的模型. 其中提升效果最好的为 VGG19 方法, 相较于单独使用预训练过后的 VGG19 模型, 在 RAF-DB 数据集中识别效果高出了 4.92%, 在 EmotioNet 数据集中高出了 4.16%, 实现了在这两个数据集下最好的识别效果, 证明了图卷

© 中国科学院软件研究所 http://www.c-s-a.org.cn

积模块可以获取表情类别之间的关系, 来更好的辅助复合人脸表情识别.



(a) RAF-DB 数据集



(b) EmotioNet 数据集

图 5 两个数据集下不同 p 值的准确率

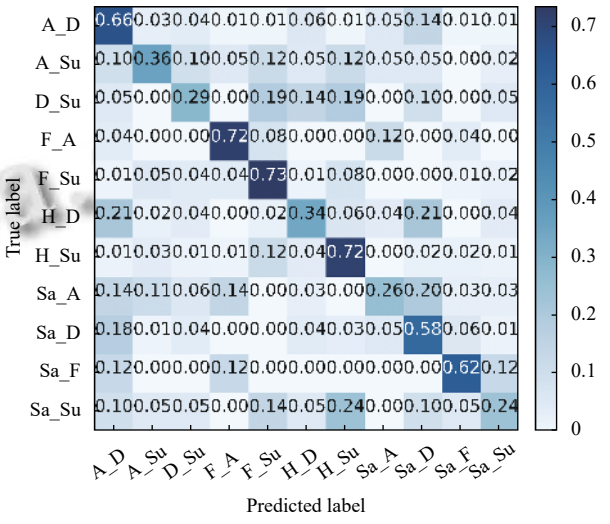
表 2 RAF-DB 数据集中 ROI 网络影响下的准确率 (%)

对比方法	准确率
GCN-ML-CFER-ROI	60.606
GCN-ML-CFER	61.995

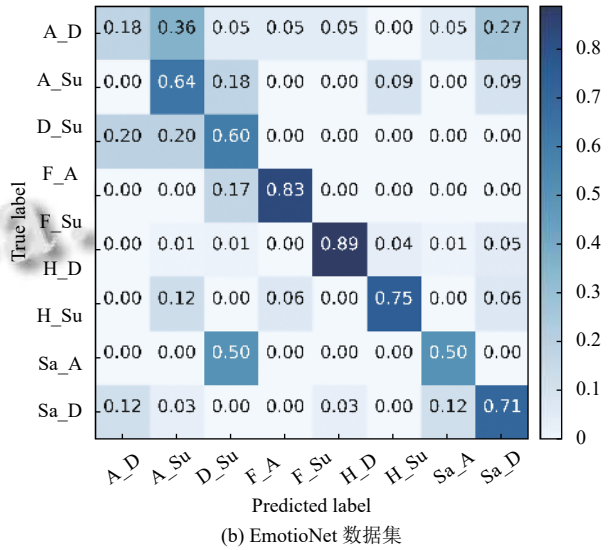
表 3 模型在不同主干网络下的准确率比较结果 (%)

GCN-ML-CGER中不同的主干网络	RAF-DB	EmotioNet
baseDCNN[9]	53.914	65.833
ResNet18[19]	59.217	77.683
ResNet34[19]	59.848	76.667
ResNet50[19]	58.864	76.250
ResNet101[19]	57.626	77.083
VGG19[19]	61.995	79.583

图 6 给出了我们模型在两个数据集上最好识别率下的混淆矩阵. 在 RAF-DB 数据集下的混淆矩阵中发现, fearfully angry 和 fearfully surprised 复合表情的识别率较高, 而 sadly angry 和 sadly surprised 复合表情的识别率较低. 在 EmotioNet 数据集下的混下矩阵中, fearfully surprised 和 happily disgusted 的识别率较高, 而 angrily disgusted 和 sadly angry 的识别率较低. 可能原因一方面在于, 数据集中复合表情的样本数目不平衡所致, 数据集中复合表情的样本数目越多, 学习到的对应的表情特征越准确, 另一方面, 构成复合表情的基本表情一起出现的概率越高, 图卷积通过语义空间学习的表情类别分类器越准确, 最后有效的提升相应复合表情的识别率.



(a) RAF-DB 数据集



(b) EmotioNet 数据集

图 6 GCN-ML-CFER 模型在两个数据集下的混淆矩阵

3 总结与展望

对于复合人脸表情识别, 本文提出了一种基于图卷积多标签学习的复合人脸表情识别方法. 针对表情类别之间的关联性, 本文将基本表情类别作为图中的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

节点, 利用先验知识和数据驱动方法, 构建了表情类别知识图, 再通过图卷积网络来有效提取知识图中的关系信息, 以提高复合人脸表情识别的性能. 在 RAF-DB 和 EmotioNet 这两个数据集上进行了大量实验, 实验结果表明: 所提出的方法在复合人脸表情识别上达到了很好的效果. 由于本文主要是对复合人脸表情进行识别, 对所有表情混合进行识别没有深入考虑, 同时, 图中节点特征的初始化使用的是词向量, 需要进一步研究是否有更合适的方式, 从而更加有效的提升人脸表情识别的准确率.

参考文献

1 Tarnowski P, Kołodziej M, Majkowski A, *et al*. Emotion recognition using facial expressions. Procedia Computer Science, 2017, 108: 1175–1184. [doi: 10.1016/j.procs.2017.05.025]

2 薛雨丽, 毛峡, 郭叶, 等. 人机交互中的人脸表情识别研究进展. 中国图象图形学报, 2009, 14(5): 764–772. [doi: 10.11834/jig.20090503]

3 张俞晴, 何宁, 魏润辰. 基于卷积神经网络融合 SIFT 特征的人脸表情识别. 计算机应用与软件, 2019, 36(11): 161–167. [doi: 10.3969/j.issn.1000-386x.2019.11.027]

4 马中启, 朱好生, 杨海仕, 等. 基于多特征融合密集残差 CNN 的人脸表情识别. 计算机应用与软件, 2019, 36(7): 197–201. [doi: 10.3969/j.issn.1000-386x.2019.07.033]

5 Li S, Deng WH. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 2018. [doi: 10.1109/TAFFC.2020.298144]

6 Kumari J, Rajesh R, Pooja KM. Facial expression recognition: A survey. Procedia Computer Science, 2015, 58: 486–491. [doi: 10.1016/j.procs.2015.08.011]

7 Du SC, Tao Y, Martinez AM. Compound facial expressions of emotion. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(15): E1454–E1462. [doi: 10.1073/pnas.1322355111]

8 Benitez-Quiroz CF, Srinivasan R, Martinez AM. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5562–5570.

9 Li S, Deng WH, Du JP. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2584–2593.

10 刘晓玲, 刘柏嵩, 王洋洋, 等. 基于深度学习的多标签生成研究进展. 计算机科学, 2020, 47(3): 192–199. [doi: 10.11896/jsjkx.190300137]

11 Wu ZH, Pan SR, Chen FW, *et al*. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4–24. [doi: 10.1109/TNNLS.2020.2978386]

12 Asif NA, Sarker Y, Chakrabortty RK, *et al*. Graph neural network: A comprehensive review on non-euclidean space. IEEE Access, 2021, 9: 60588–60606. [doi: 10.1109/ACCESS.2021.3071274]

13 Wang XL, Ye YF, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6857–6866.

14 Chen ZM, Wei XS, Wang P, *et al*. Multi-label image recognition with graph convolutional networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5172–5181.

15 Zhang MH, Liang YM, Ma HD. Context-aware affective graph reasoning for emotion recognition. 2019 IEEE International Conference on Multimedia and Expo. Shanghai: IEEE, 2019. 151–156.

16 Li GB, Zhu X, Zeng YR, *et al*. Semantic relationships guided representation learning for facial action unit recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8594–8601.

17 Sun X, Zheng SX, Fu HS. ROI-attention vectorized CNN model for static facial expression recognition. IEEE Access, 2020, 8: 7183–7194. [doi: 10.1109/ACCESS.2020.2964298]

18 Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543.

19 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

20 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

21 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

22 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations. Scottsdale: ICLR, 2013.

23 Mikolov T, Grave E, Bojanowski P, *et al*. Advances in pre-training distributed word representations. 11th International Conference on Language Resources and Evaluation. Miyazaki: ELRA, 2018. 52–55.

© 中国科学院软件研究所 http://www.c-s-a.org.cn