基于上下文感知和注意机制的多学习情绪识别方法*

万家华[1)†] 陈乃金[2)]
(1)安徽新华学院信息工程学院, 230088, 安徽合肥; 2)安徽工程大学计算机与信息学院, 241000, 安徽芜湖)

摘要 为提高人脸图像情绪识别效率与准确性,在探讨了深度神经网络、注意机制与损失函数基础上,提出基于上下文感知与注意机制的多学习情绪识别网络结构. 该网络主要由场景特征提取、身体特征提取与融合决策 3 个子网络组成,并采用单双输出结构,实现多标签情绪分类与连续空间情绪回归任务. 考虑到多标签情绪分类时标签的不平衡性,提出了一个改进的焦点损失(focal loss, FL)函数,可为小样本或难分类样本分配更多的权重,从而提高了网络训练效率. 利用 EMOTIC 数据集进行仿真,结果表明平均绝对误差回归组合损失训练性能更优,分类平均准确率与回归平均误差率分别为 28.5% 和 0.098,该方法对于小样本或难分类样本具有更好的分类效果.

关键词 人脸图像; 情绪识别; 上下文感知; 注意机制; 多标签
中图分类号 TP393 **DOI**: 10.12202/j.0476-0301.2021175

人脸图像为情绪感知提供了相关信息[1−3],因此可以通过观察人的面部特征来推断其所具有的情绪. 目前,该技术已被广泛应用于人机交互[4]、医疗护理[5−6]等领域. 在过去十几年里,卷积神经网络因在图像分类、识别等领域的优越性能,为情绪感知注入了新的活力[7]. 文献 [8] 针对公共空间个体人脸分辨率较低、表情识别精度不高的问题,提出了融合面部表情和身体姿态的情绪识别方法; 文献 [9]针对自然状态下小群体图像的情绪分类,提出基于面部、场景和骨架 3 种视觉线索的混合深度网络,独立学习不同特征,最终通过决策融合获得情绪分类; 文献 [10] 结合 Viola-Jones、自适应直方图均衡、离散小波变换与深度卷积神经网络,提出了一种面部情绪自动识别算法; 文献 [11] 提出了一种基于双流卷积神经网络算法,并在 CK+与 JAFFE 数据集上进行了验证,显示出较好的识别效果. 受自热光照条件、低分辨率成像等因素影响,室外人脸情绪识别存在诸多挑战. 传统情绪识别问题主要对 6 种情绪(中性、恐惧、惊讶、厌恶、悲伤和快乐)进行分类. 然而,只有 6 种情绪分类远不能描述人们复杂的情绪波动. 同时,情绪识别是一个重要的上下文感知环节,大部分研究集中于文字、对话等场景[12−13],因此人脸图像情绪识别仍存在巨大挑战,主要原因是由于数据的不可用性,以及用传统方法表示上下文比较困难等.

为改善传统情绪识别分类有限、上下文感知环节困难等问题,本文提出了基于上下文感知和注意机制的多学习情绪识别网络结构. 该网络主要由场景特征提取、身体特征提取与融合决策 3 个子网络组成,并采用单输入双输出结构. 特征提取网络将整个图像作为输入; 身体特征提取网络将边界框分隔的身体作为输入; 融合决策网络用于合并场景与身体特征,并输出最终决策. 本研究方法的优点有: 1)提出了一种改进的焦点损失(focal loss, FL)解决样本不平衡问题; 2)提出了一种“端到端”的基于上下文感知和注意机制的多学习情绪识别网络; 3)最终模型输出包括多标签情绪分类和连续空间情绪回归 2 部分.

1 网络结构

网络模型采用单输入双输出结构(图 1): 在输入层,选择整幅图像作为输入; 为了识别图像中人的情绪变化,必须在整体图像中找出人的细节特征.

1.1 场景特征提取网络 场景特征提取网络将包含多人的整个图像作为输入,从而提取场景全局特征. 为加快训练效率,在 Places 数据库上使用 VGG16 进行预训练. 场景特征提取网络中共包含 4 个密集特征模块,每个密集特征模块通过 2 个 1×1 卷积核来减少输入特征映射的通道数.

1.2 身体特征提取网络 身体特征提取网络以图像中隐含的面部表情、头部位置与身体姿势等信息作为输入. 与场景特征提取模块类似,通过采用在

* 国家自然科学基金资助项目(61973295); 安徽省教育厅重点科研资助项目(KJ2019A0877)
† 通信作者: 万家华(1980—),男,硕士,副教授. 研究方向: 数据挖掘、机器学习. E-mail: wanjiahua2009@163.com
收稿日期: 2021-07-22

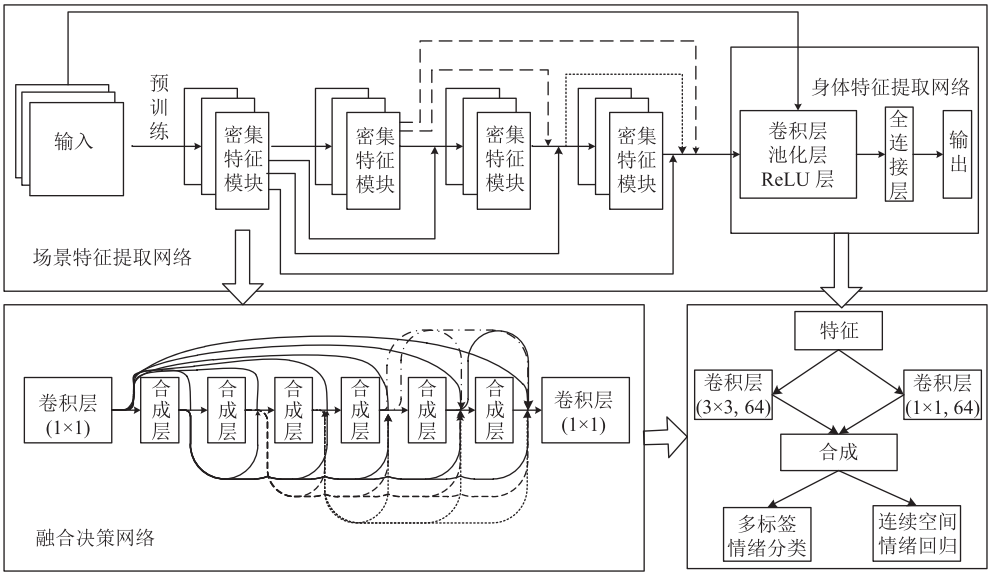

图1 基于上下文感知和注意机制的多学习情绪识别网络结构示意

imagenet上预先训练网络,从而加快训练效率.身体特征提取网络由5个卷积层和1个全连接层组成的"端到端"结构构成.为了提高网络的非线性度,在每个卷积层后接批量归一化(batch normalization, BN)层、池化层与ReLU层.池化层可以防止特征图的过拟合,降低特征图的维数,且更好地保留输入特征图的全局信息.

1.3 融合决策网络 将场景特征与身体特征网络输出的特征信息进行重构,并代入融合决策网络.该网络包含1×1卷积核与3×3卷积核并行执行卷积运算,并拼接所有输出结果.不同的卷积运算可以获得输入图像的不同信息,而并行运算后的特征映射表现出更强的特征表示能力.进一步,利用6个合成层建立不同层间的直接连接,充分利用各层的特征映射,并结合各通道的特征,以缓解梯度消失问题.同时,在每一层中使用BN层与ReLU函数,这样可防止梯度消失,增加了网络的非线性程度.模型输出包括多标签情绪分类与连续空间情绪回归2部分.

2 网络运行

由于模型中多标签情绪分类是1个分类任务,连续空间情绪回归为1个回归任务,因此2个任务分别采用了不同的损失函数.

全局损失函数为2个不同损失函数的加权和,即

$$F=\lambda_1 F_{\mathrm{ML}}+\lambda_2 F_{\mathrm{G}}, \tag{1}$$

式中:F_{ML}与F_{G}分别表示多标签分类和连续变量回归损失函数;λ_1和λ_2为不同损失权重.本文采用了几种不同的损失函数组合方案.

2.1 多标签分类损失 考虑到多标签情绪分类时标签的不平衡性,本文提出了一种改进的FL函数.

FL函数[14]是目标检测领域处理样本不平衡问题时常用手段之一.该函数的机制为小样本或难分类样本分配更多的权重,从而加强了对纠正错误分类示例的重视.一般情况下,FL函数定义如下:

$$F_{\mathrm{L}}(\sigma_t)=-\alpha(1-\sigma_t)^{\gamma}\log_n(\sigma_t), \tag{2}$$

$$\sigma_t=\begin{cases}\sigma, & y=1,\\ 1-\sigma, & \text{其他},\end{cases} \tag{3}$$

式中:γ为焦点参数;σ和y分别是正类样本的分类模型输出与它的真实标签.

本文的情绪分类不仅存在样本不平衡问题,还是一个多标签分类问题.因此,为提高多标签情绪分类效率,提出改进的FL(多标签分类)函数(FML),即

$$F_{\mathrm{ML}}(y,\sigma)=-\sum_{i=1}^{M_c}\alpha(1-\sigma_i)^{\gamma}y_i\log_n(\sigma_i)+(1-\alpha)\sigma_i^{\gamma}(1-y_i)\log_n(1-\sigma_i), \tag{4}$$

式中:M_c是分类的数量,$M_c=26$;σ_i和y_i分别表示分类模型输出与它的第i个真实标签;α与γ为超参数,α为召回率或精度所占比例,γ为焦点参数,与传统焦点损失参数相同,γ的主要作用是增加难分类样本的权重,降低易分类样本的权重,从而提升整体样本平衡性.

此外,为了比较损失函数的运行性能,本文还对交叉熵、欧式距离等分类损失函数,以及Huber损失、均方差损失、平均绝对误差等回归损失进行了对比分析.

分类中交叉熵和欧式距离损失函数F_{C}与F_{E}分别定义为

$$F_{\mathrm{C}}(y,\sigma)=-\sum_{i=1}^{M_c}y_i\log_n(\sigma_i)+(1-y_i)\log_n(1-\sigma_i), \tag{5}$$

$$F_{\mathrm{E}}(y,\sigma)=\sum_{i=1}^{M_{\mathrm{c}}}(\sigma_i-y_i)^2, \tag{6}$$

式中σ_i、y_i、M_{c}与F_{ML}中所给的参数定义相同.

2.2　连续变量回归损失　回归问题中, 本文采用了均方差损失$F_{\mathrm{G}}^{\mathrm{L}_2}$与平均绝对误差$F_{\mathrm{G}}^{\mathrm{L}_1}$2 种不同测量法, 分别定义为

$$F_{\mathrm{G}}^{\mathrm{L}_2}(y,\sigma)=\frac{1}{M_{\mathrm{G}}}\sum_{i=1}^{M_{\mathrm{G}}}(\sigma_i-y_i)^2, \tag{7}$$

$$F_{\mathrm{G}}^{\mathrm{L}_1}(y,\sigma)=\frac{1}{M_{\mathrm{G}}}\sum_{i=1}^{M_{\mathrm{G}}}|\sigma_i-y_i|, \tag{8}$$

式中M_{G}为连续变量维度, $M_{\mathrm{G}}=3$.

3　仿真与分析

3.1　仿真环境与数据集　仿真软件为 Python 语言(PyTorch1.5.1+CUDA10.0).

仿真所用数据集为 EMOTIC 数据集. EMOTIC 数据集收集了人们在不同地方进行不同活动时所表现出的各种各样情绪状态的多组图片与注释. 该数据集有 18316 张图片、23788 个注释. 注释可分为离散和连续维度 2 类: 离散类共包含 26 种情绪类别(表 1); 连续维度为价-唤醒-优势(valence-arousal-dominance, VAD)三维情绪状态模型构成的数据集, 又进一步由正面到负面情绪映射到 1~10 的得分空间(表 2).

表 1　离散情绪分类

情绪	占比/%	情绪	占比/%
平和	7.4	喜爱	4.9
期待	23.1	投入	55.1
开心	25.0	愉悦	9.4
惊讶	1.8	同情	3.3
尊敬	3.8	孤立	6.5
自信	17.9	疲劳	2.3
激动	19.6	尴尬	0.6
困惑	2.9	渴望	2.7
反对	1.5	厌恶	0.8
恼怒	1.7	愤怒	0.9
敏感	1.6	伤心	1.8
不安	2.2	恐惧	0.8
苦恼	0.8	煎熬	1.7

3.2　损失性能对比　系统训练时网络超参数具体有: 学习率(10^{-4})和学习率衰减倍数(10^{-2}), 学习率衰减周期(15 代)和最大训练次数(150 代), 批处理(16)和 Dropout 率(0.2), γ(3.0)、λ_1(0.5)及 λ_2(0.5)等. 将数据集分为训练集、测试集和验证集, 比例为 7∶1∶2. 代入基于上下文感知和注意机制的多学习情绪识别网络模型. 对不同组合损失下模型训练效果进行对比分析.

表 2　连续情绪不同维度空间分数分布

得分	价维度占比/%	唤醒维度占比/%	优势维度占比/%
1	0.3	0.7	0.5
2	0.7	5.1	1.7
3	2.3	7.9	4.7
4	7.2	20.6	6.5
5	21.3	13.7	8.2
6	35.5	15.9	22.9
7	18.1	17.7	28.7
8	11.8	9.1	12.3
9	1.7	4.4	9.4
10	0.7	3.3	3.5

3.2.1　分类损失性能对比　图 2 是不同损失组合下多标签分类训练与验证平均准确率对比曲线, 蓝色线表示由欧式距离分类与平均绝对误差回归损失共同组合成的最终损失训练效果. 从图 2 中可以看出: 该组合下 80 代左右训练趋于稳定, 且平均准确率最终收敛于 0.281; 橙色线表示由欧式距离分类损失和均方差损失组合成的损失训练效果, 该组合下约 80 代训练趋于稳定, 且平均准确率最终收敛于 0.283; 绿色线表示所提改进焦点损失和平均绝对误差回归损失组合成损失训练效果, 该组合下约 20 代训练趋于稳定, 且平均准确率最终收敛于 0.285; 红色线表示由改进焦点和均方差损失组合成的损失训练效果, 该组合下约 110 代训练趋于稳定, 且平均准确率最终收敛于 0.280; 紫色线表示由交叉熵损失和均方差损失组合成的损失训练效果, 该组合下约 120 代训练趋于稳定, 且平均准确率最终收敛于 0.272; 棕色线表示由交叉熵损失和平均绝对误差回归损失组合成的损失训练效果, 该组合下约 110 代训练趋于稳定, 且平均准确率最终收敛于 0.274.

3.2.2　回归损失性能对比　图 3 为不同损失组合下连续回归平均误差率对比曲线, 由图 3 可知, 蓝色线、橙色线、绿色线、红色线、紫色线及棕色线组合下平均误差率最终分别收敛于 0.122、0.115、0.098、0.119、0.129 及 0.123.

由综合分析可知: 改进焦点损失与平均绝对误差回归组合损失时训练性能更优, 该组合下训练速度快, 且分类平均准确率与回归平均误差率分别为 28.5% 和 0.098, 明显优于其他方案.

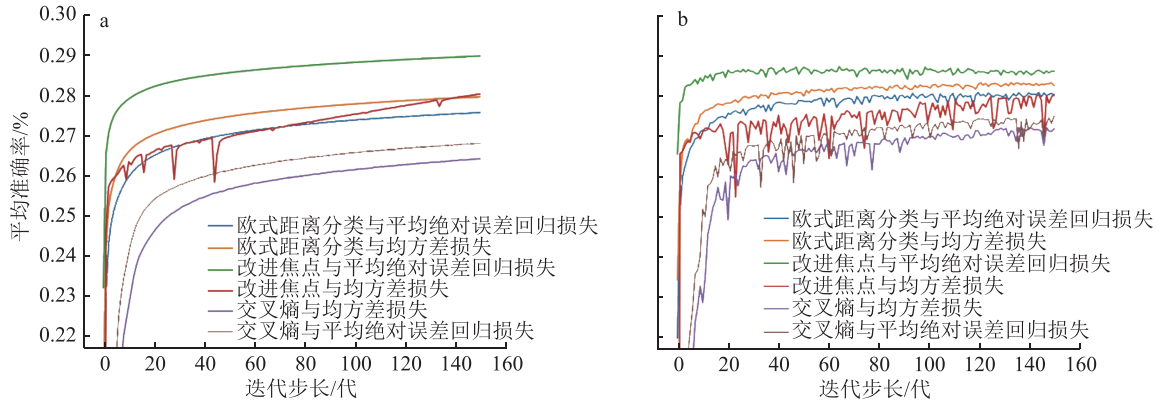

图 2 不同组合下训练（a）、验证（b）平均准确率变化对比结果

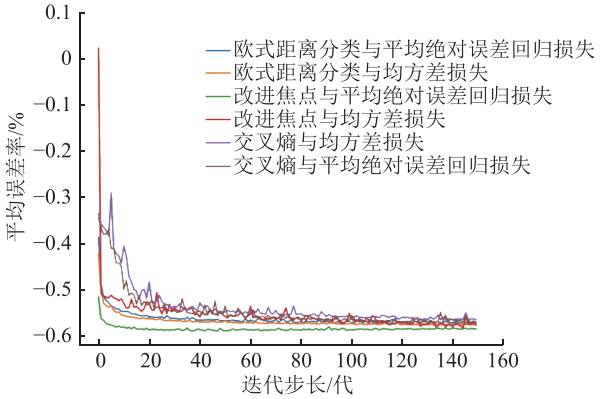

图 3 回归损失组合下平均误差率对比结果
验证损失变化曲线

3.3 不同方法对比分析 将本文所提出的新方法与文献 [15−16] 所提出的方法在 EMOTIC 数据集上的运行性能进行了对比，结果如表 3 所示. 从表 3 中可以看出：文献 [15] 效果最低，平均准确率仅为 27.4%；文献 [16] 与本文所提方法差距不大，然而对于占比<3%的分类样本，本文提出的模型具有更好的分类效果. 原因如下：1）改进 FL 为小样本或难训练样本能提供更多焦点；2）本文方法是基于注意力机制，这使得网络更加关注细节特征，因此，本文方法平均分类准确率更高.

表 3 不同模型平均准确率对比结果

情绪	平均准确率/%		
	本文方法	文献[15]	文献[16]
平和	24.3	21.6	38.9
期待	57.7	58.6	62.6
开心	76.6	58.3	73.3
惊讶	10.1	18.8	9.0
尊敬	16.7	17.7	13.3
自信	75.3	78.4	72.5
激动	70.4	77.2	71.9
困惑	23.1	29.6	18.7
反对	14.9	15.0	11.3
恼怒	17.4	14.1	11.3
敏感	6.3	9.3	6.1
不安	19.7	16.9	16.5
苦恼	9.4	8.8	6.6
喜爱	32.0	27.9	57.1
投入	85.9	87.7	88.3
愉悦	47.0	46.1	58.4
同情	14.1	14.7	18.1
孤立	28.3	22.5	27.7
疲劳	15.3	9.2	12.3
尴尬	2.9	3.2	1.2
渴望	9.9	8.2	10.3
厌恶	9.3	7.9	5.9
愤怒	14.3	9.5	10.2
伤心	24.1	19.7	25.6
恐惧	9.2	13.2	4.1
煎熬	28.6	19.5	9.8
平均	28.6	27.4	28.5

4 结论

利用深度学习技术对多标签情绪分类进行了分析与探索，提出了一种“端到端”的、基于上下文感知和注意机制的多学习情绪识别网络，并在 EMOTIC 数据集中进行了验证. 为有效处理不平衡数据，提出了一个改进的焦点损失函数，为小样本或难分类样本分配更多的权重. 对不同组合损失下网络性能进行了对比，结果表明：本文提出的改进焦点与平均绝对误差回归组合损失使网络性能更优. 未来工作一方面将着重提取情感状态特征，关注情感变化梯度，从而按时序对复杂情绪进行分割与细类识别；另一方面，将本模型与机器学习模型如粒子群、神经网络等优化算法结合，计算出实际问题中的最优阈值，从而进

一步提高训练效率.

5　参考文献

[1] 刘智, 方常丽, 刘三妍, 等. 物理学习空间中学习者情绪感知研究综述[J]. 远程教育杂志, 2019, 37(2): 33

[2] 李婷婷, 胡玉龙, 魏枫林. 基于GAN改进的人脸表情识别算法及应用[J]. 吉林大学学报(理学版), 2020, 58(3): 605

[3] 党宏社, 王淼, 张选德. 基于深度学习的面部表情识别方法综述[J]. 科学技术与工程, 2020, 20(24): 9724

[4] 张凯乐, 刘婷婷, 刘箴, 等. 面向情绪调节的多模态人机交互技术[J]. 中国图象图形学报, 2020, 25(11): 2451

[5] 张薛晴, 程立辉, 宋玉磊, 等. 情感识别技术在轻度认知障碍中的应用现状[J]. 护理研究, 2020, 34(10): 1750

[6] 赵玲玲, 裴炬盛, 韦荣泉, 等. 癌症患者医疗风险感知的研究进展[J]. 医学与哲学, 2021, 42(6): 45

[7] 潘家辉, 何志鹏, 李自娜, 等. 多模态情绪识别研究综述[J]. 智能系统学报, 2020, 15(4): 633

[8] 文虹茜, 卿粼波, 晋儒龙, 等. 基于表情及姿态融合的情绪识别[J]. 四川大学学报(自然科学版), 2021, 58(4): 87

[9] 季欣欣, 邵洁, 钱勇生. 基于注意力机制和混合网络的小群体情绪识别[J]. 计算机工程与设计, 2020, 41(6): 1683

[10] 王春峰, 李军. 基于面部检测和深度神经网络的面部情绪自动识别算法[J]. 光电子·激光, 2020, 31(11): 1197

[11] 翟海庆, 刘丹, 刘畯. 利用双流卷积神经网络的人脸表情识别方法[J]. 光学技术, 2020, 46(6): 712

[12] 张晨, 钱涛, 姬东鸿. 基于神经网络的微博情绪识别与诱因抽取联合模型[J]. 计算机应用, 2018, 38(9): 2464

[13] 赖河蒗, 李玲俐, 胡婉玲, 等. 一种层次化R-GCN的会话情绪识别方法[J/OL]. (2021-02-07)[2021-07-22]. https://doi.org/10.19678/j.issn.1000-3428.0060346

[14] 张凯琳, 阎庆, 夏懿, 等. 基于焦点损失的半监督高光谱图像分类[J]. 计算机应用, 2020, 40(4): 1030

[15] KOSTI R, ALVAREZ J M, RECASENS A, et al. Context based emotion recognition using EMOTIC dataset[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(11): 2755

[16] ZHANG M H, LIANG Y M, MA H D. Context-aware affective graph reasoning for emotion recognition[C]//2019 IEEE International Conference on Multimedia and Expo (ICME), July 8−12, 2019, Shanghai International Convention Center. 2019: 151

Multi learning emotion recognition based on context awareness and attention mechanism

WAN Jiahua[1)†]　CHEN Naijin[2)]

(1)School of Information Engineering, Anhui Xinhua University, 230088, Hefei, Anhui, China;
2)School of Computer and Information, Anhui Engineering University, 241000, Wuhu, Anhui, China)

Abstract　To improve the efficiency and accuracy of facial image emotion recognition, a multi learning emotion recognition network structure was proposed based on context awareness and attention mechanism. The proposed network was composed of three sub networks: scene feature extraction, body feature extraction and fusion decision-making. Single and double output structures were adopted to realize multi-label emotion classification and continuous spatial emotion regression. Improved focus loss function was proposed to assign more weights to small samples or samples that were difficult to classify, to improve efficiency of network training. Simulations using emotic data set showed that proposed improved focus loss and mean absolute error regression combination loss was better, average classification accuracy and regression average error rate were 28.5% and 0.098 respectively. It is concluded that the proposed method had better classification effect for small samples.

Keywords　face image; emotion recognition; context awareness; attention mechanism; multi-label

【责任编辑: 陆有忠】