计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-3029-10
doi: 10.3778/j.issn.1673-9418.2210079

开放科学(OSID)

类别平衡调制的人脸表情识别

刘成广[1,2],王善敏[3],刘青山[1,2+]
1. 南京信息工程大学 计算机学院,南京 210044
2. 南京信息工程大学 数字取证教育部工程研究中心,南京 210044
3. 南京航空航天大学 计算机科学与技术学院,南京 211106
+ 通信作者 E-mail: qsliu@nuist.edu.cn

摘　要:人脸表情识别(FER)旨在从人脸图片中判断表情的类别,在心理诊断、人机交互等领域有着广泛的应用前景。在实际任务中,不同表情数据的分布往往是不平衡的。数据的不平衡导致了各表情的特征分布不平衡和分类器优化不平衡,从而影响了表情识别模型的性能。为此,提出了一种类别平衡调制的人脸表情识别方法(CBM-Net),分别在特征学习阶段和分类器优化阶段对模型进行类别平衡调制。CBM-Net包括特征调制和梯度调制两个模块。特征调制模块通过在特征方向上增加类间的可分性与类内的紧密性,实现各类别的特征分布平衡。梯度调制模块利用批次训练样本的统计信息对各分类器的优化梯度进行反向调节,确保各类别的分类器收敛速度一致,使得各分类器性能同时达到最优。在四个流行的数据集上进行的定性和定量实验表明,CBM-Net在人脸表情识别的类别平衡调制上是有效的,与一众先进方法相比,效果也相当良好。

关键词:人脸表情识别(FER);类别不平衡;类别平衡调制;特征调制;梯度调制
文献标志码:A　　**中图分类号:**TP391.4

Class-Balanced Modulation for Facial Expression Recognition

LIU Chengguang[1,2], WANG Shanmin[3], LIU Qingshan[1,2+]
1. School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China
2. Engineering Research Center of Digital Forensics Ministry of Education, Nanjing University of Information Science & Technology, Nanjing 210044, China
3. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract: Facial expression recognition (FER) aims at determining the types of facial expressions for given facial images, which has a broad application prospect in psychological diagnosis, human-computer interaction, etc. In practical tasks, various databases tend to have imbalanced data distributions among basic facial expressions. Such an issue has caused imbalanced feature distribution and inconsistent classifier optimization for various facial expressions, seriously affecting the performance of expression recognition models. To solve this issue, this paper proposes a class-balanced modulation mechanism for facial expression recognition (CBM-Net), which attempts to address the imbalanced data distribution problem by modulating the FER model in feature learning and classifier optimization stages. CBM-Net includes two modules of feature modulation and gradient modulation. The feature modulation module struggles to balance feature distributions for all facial expressions by increasing the separability

基金项目:江苏省自然科学基金(BK20192004B)。
This work was supported by the Natural Science Foundation of Jiangsu Province (BK20192004B).
收稿日期:2022-10-19　　**修回日期:**2023-01-04

between classes and the tightness within classes in the feature direction. The gradient modulation module uses the statistical information of batch training samples to reversely adjust the optimization gradient of each classifier to ensure that the convergence speed of each classifier is consistent, so that the performance of each classifier can be optimal at the same time. Qualitative and quantitative experiments on four popular datasets show that CBM-Net is effective in class-balanced modulation, and its effect is quite good compared with many advanced methods.

Key words: facial expression recognition (FER); class imbalance; class balance modulation; feature modulation; gradient modulation

面部表情是人类传达情感信息的最直接方式[1-2]。自动的人脸表情识别(facial expression recognition,FER)有着广泛的应用,如人机交互[3]、心理健康问诊[4]、疲劳驾驶检测[5]等。近年来,由于大型数据集的出现,基于深度学习的表情识别技术取得了较大进展。然而,与实验室环境下采集的数据集(例如CK+[6]、MMI[7]和JAFFE[8])相比,自然场景下采集的数据集存在明显的类别不平衡问题[9-11]。图1展示了RAF-DB数据集各类别样本的分布和对应的识别精度。数据分布不平衡导致了模型对各类表情的识别精度差异较大。具体地,模型对样本量大的类别识别精度较高,而对样本量少的类别识别精度较低。

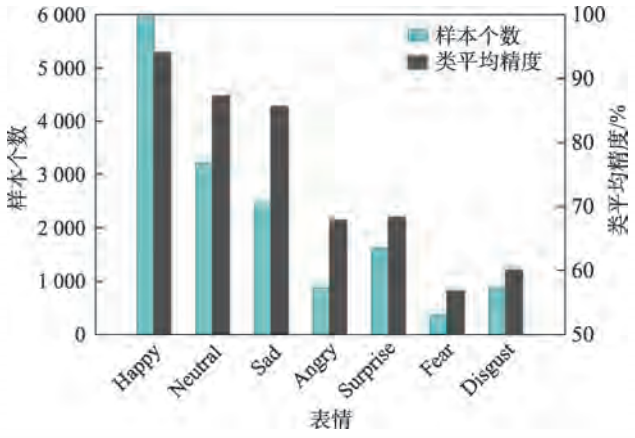

图1 类别不平衡示例图

Fig.1 Example of class imbalance

为了解决类别不平衡问题,常用的方案通常统计数据集中各类别的样本分布,对样本进行重采样(re-sampling)[12-15]或重加权(re-weighting)[16-20]。重采样的方式通过数据增强对少样本类别进行上采样,或者对多样本类别进行随机删除的下采样。然而上采样没有给模型带来更多实质性的信息,下采样明显减少了模型的训练数据。因此,重采样的方法没有从本质上解决表情识别任务中样本不平衡的问题。重加权的方法依据类别样本数或样本容量[21],对样本反向加权,强化少样本类别的学习[16-17]。但是,研究表明,图像的特征分布和类别的标注分布是不耦合的[22-23],用类别的分布影响表征学习时的特征分布是不合适的。因此,将图像表征学习和分类器优化过程联合调制的重加权方法,不仅没有平衡各类别的特征分布,而且会影响图像表征的正常学习过程。

为此,本文将图像表征学习和分类器优化过程分离,提出了一种新的类别平衡调制的人脸表情识别方法(class-balanced modulation mechanism for facial expression recognition,CBM-Net),以解决数据不平衡导致的特征分布不平衡和分类器优化不平衡问题。具体地,分别设计了特征调制和梯度调制两个模块。特征调制模块通过增加类间的方向可分性,进而确保模型可以提取出小类样本的区分性特征,使得不同类别在特征分布上保持平衡。梯度调制模块利用每批次训练样本的统计信息来调节各类别分类器的梯度,使得样本数较少的类获得更多优化,获得足够的训练尝试,而不影响其他的类。为了验证该方法的有效性,本文在RAF-DB[9]、AffectNet[10]、SFEW[24]和CAER-S[25]四个流行的数据集上进行了实验。定性和定量结果都证明了CBM-Net在解决类别分布不平衡问题上的合理性和优越性。

本文工作的主要贡献总结如下:

(1)提出了一种类别平衡调制的人脸表情识别方法,该方法从特征调制和分类器梯度调制两方面解决类别不平衡问题。

(2)设计了一个特征调制模块来保证特征间的类别可分性,进而解决类别不平衡导致的表情特征分布不平衡的问题。

(3)设计了一个梯度调制模块对分类器的优化过程进行调制,进而解决分类器中收敛速度不一致的问题。

1 相关工作

对类别不平衡问题的现有解决方案可以被分为两类：数据层面的重采样[12-15]和算法层面的权重分配[16-20]。

1.1 重采样

重采样的方法分为上采样和下采样两种。上采样主要是对较小的类进行数据增强，以获得更多的样本[12-13]。然而，上采样方式虽然增加了较小类的样本数，但是完全依赖于数据增强的方式。通过旋转、剪切、平移等简单数据增强式获得的上采样样本与原数据可能高度相似，对模型的训练并没有本质的性能提升[21]。通过生成模型（generative adversarial networks，GAN）来实现数据增强的方式，较难生成细粒度的表情图像。下采样主要是从较大的类中随机选择较少的样本用于训练，以平衡不同类的样本量。然而，这种方式可能使得参与学习的有用信息减少，最终影响模型的学习性能。

1.2 重加权

重加权的方法常常为小类样本在损失函数上赋予高权重，使其获得更多的优化。各类别权重的计算方式主要分为统计样本量占比[16-17]以及评估样本难度两种方式[18-20]。目前，统计样本量占比的方法获得了广泛的应用。通常，分配的权重与该类样本数成反比。此外，评估样本难度的方法认为来自小类的样本往往比来自大类的样本更难学习，因为小类的样本的表征学习更差。更难学习即预测损失函数值很大，将损失值作为样本的权重[20]。部分工作[26-27]认为样本难度与样本数量之间没有直接关系。由于图像的特征分布和类别的标注分布不耦合的本质[22]，利用分配权重的损失函数将影响模型对图像的表征学习，并没有为图像表征过程带来更多的提升，各类别的特征分布仍然是不平衡的。为此，采用两阶段，将图像表征阶段与分类器调制阶段分离的平衡调制方式值得考虑。

2 本文方法

2.1 不平衡问题分析

人脸表情识别旨在对人脸图片提取表情特征，并推断表情的类别。具体来说，给定训练数据集为 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$。其中，$\boldsymbol{x}_i$，$y_i$ 分别是训练样本和标签。$y_i\in\{1,2,\cdots,C\}$，C 为表情类别数。N 为样本总数。针对训练样本 $(\boldsymbol{x}_i,y_i)$，首先，使用CNN的骨干网络 $\varphi(\cdot)$ 提取图像的特征，提取过程表示如下：

$$\boldsymbol{f}_i=\varphi(\boldsymbol{\theta},\boldsymbol{x}_i) \tag{1}$$

其中，$\boldsymbol{\theta}$ 为骨干网络参数。$\boldsymbol{f}_i\in\mathbf{R}^M$，$M$ 为特征维度。当类别不平衡时，由于小类样本数量较少，在特征质量上也难以有与大类明显区分的关键性特征，致使小类样本的特征与大类样本特征区分不明显，分布不平衡，不利于后续分类。

在获得图像的特征 $\boldsymbol{f}_i$ 后，使用线性分类器建立特征 $\boldsymbol{f}_i$ 到各类别预测概率 $\boldsymbol{p}$ 的映射，取概率值最大的类别作为样本的预测表情 y'。各类别预测概率 $\boldsymbol{p}$ 计算如下：

$$\boldsymbol{p}(\boldsymbol{x}_i)=\boldsymbol{W}(\boldsymbol{f}_i)+\boldsymbol{b} \tag{2}$$

其中，$\boldsymbol{p}(\boldsymbol{x}_i)\in\mathbf{R}^C$。$\boldsymbol{W}$ 为分类器最后一层线性参数，$\boldsymbol{W}\in\mathbf{R}^{M\times C}$。$\boldsymbol{b}$ 为偏置项，$\boldsymbol{b}\in\mathbf{R}^C$。预测表情 $y'=\operatorname{argmax}(\boldsymbol{p})$。根据输入特征与各类别预测概率的对应关系，分类器 $\boldsymbol{W}$ 可进一步表示为 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_C]$。设该样本 $\boldsymbol{x}_i$ 属于第 c 类，将第 c 类的逻辑输出表示为 $\boldsymbol{p}(\boldsymbol{x}_i)_c$，则 $\boldsymbol{x}_i$ 的损失为 $L=-\mathrm{lb}(\boldsymbol{p}(\boldsymbol{x}_i)_c)=-\mathrm{lb}(\boldsymbol{f}_i\cdot\boldsymbol{w}_c)$。当模型根据样本 $\boldsymbol{x}_i$ 计算的损失函数优化分类器参数时，$\boldsymbol{W}$ 的参数更新如下：

$$\boldsymbol{W}^{t+1}=\boldsymbol{W}^t-\eta\cdot\nabla_{\boldsymbol{W}}L(\boldsymbol{W}^t) \tag{3}$$

其中，η 为学习率。由于 $\boldsymbol{x}_i$ 的标签为 y_i 属于第 c 类，损失反传 $\boldsymbol{W}$ 梯度为：

$$\nabla_{\boldsymbol{W}^t}L(\boldsymbol{W}^t)=\left[0,\cdots,\frac{\partial L(\boldsymbol{f}_i\cdot\boldsymbol{w}_c)}{\partial\boldsymbol{f}_i},\cdots,0\right] \tag{4}$$

即当样本 $\boldsymbol{x}_i$ 属于第 c 类时，损失反传仅更新第 c 类的权重 $\boldsymbol{w}_c$。设训练集中各类样本数为 $[N_1,N_2,\cdots,N_c,\cdots,N_C]$，其中 $\sum_{c=1}^{C}N_c=N$，则交叉熵损失函数为：

$$L_{\mathrm{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\mathrm{lb}\frac{\mathrm{e}^{\boldsymbol{p}(\boldsymbol{x}_i)_{y_i}}}{\sum_{j=1}^{C}\mathrm{e}^{\boldsymbol{p}(\boldsymbol{x}_i)_j}}=\sum_{c=1}^{C}\frac{N_c}{N}\boldsymbol{P}_c \tag{5}$$

其中，$\boldsymbol{P}_c=-\frac{1}{N_c}\sum_{i=1}^{N_c}\mathrm{lb}\frac{\mathrm{e}^{\boldsymbol{p}(\boldsymbol{x}_i')_{y_i=c}}}{\sum_{j=1}^{C}\mathrm{e}^{\boldsymbol{p}(\boldsymbol{x}_i')_j}}$。由此，第 c 类的梯度为 $\frac{\partial L_{\mathrm{CE}}}{\partial\boldsymbol{w}_c}=\frac{N_c}{N}\frac{\partial\boldsymbol{P}_c}{\partial\boldsymbol{w}_c}$，即梯度与类样本数占总样本数比重呈现正相关。而各类样本的损失反传仅更新当前类权重，当类别不平衡时，将导致 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_C]$ 的优化不平衡，分类器收敛速率不协调。当多样本类别达到收敛饱和时，少样本类别可能仍然未完全收敛，需要进一步训练。当少样本类别达到收敛饱和时，多样本类别可能已经过拟合。

2.2 方法概述

为了解决类别不平衡导致的特征分布不平衡和分类器优化不平衡的问题，本文提出了一种类别平衡调制的人脸表情识别方法（CBM-Net）。如图2所示，CBM-Net包括常规的深度学习骨干网络以及特征调制和梯度调制两个分支模块。CBM-Net输入为一批人脸表情图像 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$，输出为对应的预测表情 y'。为了应对数据不平衡情况，本文将图像的表征阶段与分类器调制阶段分离。具体地，在反向的调制过程中，分别应用特征调制和梯度调制模块优化特征分布和分类器参数。首先，针对骨干网络提取的样本特征 $\boldsymbol{f}_i$，通过最大化类间的方向性，进而确保模型可以提取出小类样本的区分性特征，使得不同类在特征分布上保持平衡。其次，针对输出概率 $\boldsymbol{p}$，计算其与标签 y 的交叉熵获得分类器梯度，利用批次样本统计信息 k 来调节分类器梯度，使得欠优化的小类获得更多的优化。

2.3 特征调制模块

特征调制模块作用于图像表征阶段，通过约束类间特征的方向性，进而确保样本不平衡的类在特征分布上保持平衡。该模块的核心是特征调制损失 L_{FM}。受文献[28]启发，L_{FM} 依据特征的相似性，增加类间距离的同时，减少了类内距离。

对于给定的两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，其特征分别为 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$，通过余弦相似度计算它们之间的特征相似性，即：

$$cs(\boldsymbol{x}_i,\boldsymbol{x}_j)=cs(\boldsymbol{f}_i,\boldsymbol{f}_j)=\frac{<\boldsymbol{f}_i,\boldsymbol{f}_j>}{\|\boldsymbol{f}_i\|\|\boldsymbol{f}_j\|} \tag{6}$$

特征调制损失 L_{FM} 可以表示如下：

$$L_{\mathrm{FM}}=1+\frac{1}{2N}\sum_{i=1}^{N}\left(\frac{1}{N_{\tilde{y}_i}}\sum_{j=1,y_j\neq y_i}^{N_{\tilde{y}_i}}cs(\boldsymbol{f}_i,\boldsymbol{f}_j)-\frac{1}{N_{y_i}}\sum_{j=1,y_j=y_i}^{N_{y_i}}cs(\boldsymbol{f}_i,\boldsymbol{f}_j)\right) \tag{7}$$

其中，N_{y_i} 是属于 y_i 类的样本数，$N_{\tilde{y}_i}$ 是剩余类的样本数，满足 $N_{y_i}+N_{\tilde{y}_i}=N$，$i,j$ 是索引。

2.4 梯度调制模块

如前所述，模型的优化过程通常由大类主导控制，从而小类的性能未能充分优化。为了解决该问题，梯度调制模块作用于分类器调制阶段，其利用批次样本的统计信息来调节各类别梯度，使得小类获得更多的优化。为简单理解，以第 c 类为例。在第 t 批次中，定义 N_c^t 为该批次第 c 类样本数，$c\in\{1,2,\cdots,C\}$，C 是表情类别数。N^t 为该批次样本总数。定义样本比 r_c^t：

$$r_c^t=\frac{N^t}{CN_c^t} \tag{8}$$

即样本数越少，该类在当前批次将获得更多优化。为自适应调节梯度，设计梯度权重 k_c^t：

$$k_c^t=m(r_c^t)=\begin{cases}1+\tanh(a\cdot r_c^t), & r_c^t>1\\ 1, & r_c^t\leqslant 1\end{cases} \tag{9}$$

其中，a 是控制调制程度的超参数。将系数 k_c^t 整合到SGD（stochastic gradient descent）优化方法中，更新分类器第 c 类梯度，更新如下：

$$\boldsymbol{w}_c^{t+1}=\boldsymbol{w}_c^t-\eta\cdot k_c^t\cdot\tilde{g}(\boldsymbol{w}_c^t) \tag{10}$$

其中，$\tilde{g}(\cdot)$ 为梯度 $\nabla_{w_c^t}L(\boldsymbol{w}_c^t)$ 无偏参数估计。依此，样本量更少的类（$r_c^t>1$）可以获得更多的优化，获得足够的训练尝试，而不影响其他的类。

2.5 损失函数

在训练过程中，CBM-Net有两个约束函数，分别

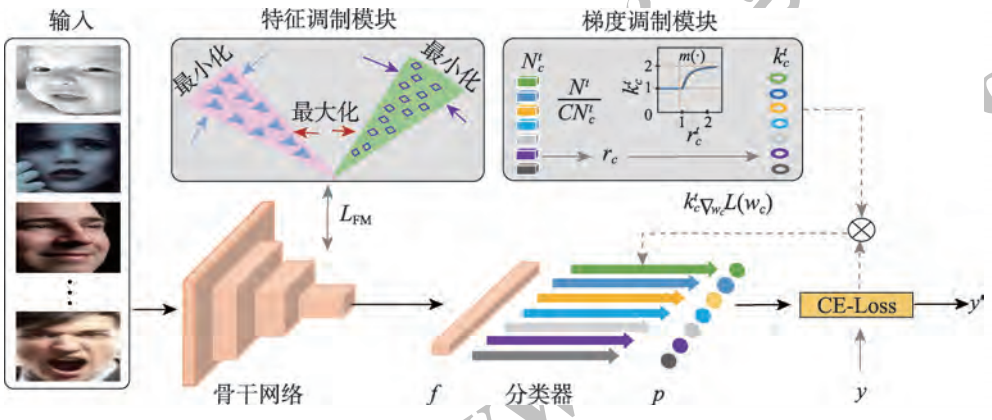

图2 网络框架图

Fig.2 Network framework diagram

是交叉熵损失函数 L_{CE} 以及特征调制函数 L_{FM}。

综上，本文所采用的损失函数为：

$$L_{SUM} = L_{CE} + \lambda L_{FM} \tag{11}$$

λ 为平衡超参数。

3 实验分析

3.1 实验数据

RAF-DB[9]：包含了30 000张带有基本或复合注释的表情图片。在实验中，为了对比公平，本文仅使用数据集中的7种基本表情，包括6种离散的基本表情和1种中性表情。

AffectNet[10]：从搜索引擎中查询与情感相关关键词收集的100多万张图像，是目前公开可用的最大FER数据集。其中，超过44万张图像被手动标注为8种表情(7种基本表情+蔑视)。本文使用28万个训练样本和4万个测试样本，分别在7种基本表情和包括蔑视的8种表情上进行实验。

SFEW[24]：采集于同一个电影的静态图片，并被标注为了7种基本情绪。被划分为了958张训练图片和436张测试图片。

CAER-S[25]：基于CAER选择的视频静态帧获得的数据集。该数据集被独立注释为7类基本表情，包含65 983张图像，其中44 996张图像用于训练，20 987张图像用于测试。

3.2 实验设计

(1)预处理和面部特征。在CBM-Net中，图像通过RetinaFace[29]进行人脸检测和对齐，并通过数据预处理进一步调整为224×224像素。CBM-Net由Pytorch实现，主干网为ResNet18[30]，从主干网最后一个池化层提取512维特征。随后，特征由512维降低到10维。数据预处理过程包括基础性数据增强和标准化。在与各种最先进的方法进行比较时，主干网ResNet-18在MS-Celeb-1M[31]人脸识别数据集上进行预训练。

(2)训练。本文使用1个Nvidia Titan 2080s GPU以端到端的方式训练CBM-Net，并将批量大小设置为128。整个网络采用 L_{FM} 和 L_{CE} 联合优化。

3.3 消融实验

3.3.1 评估CBM-Net的不同模块

为验证本文提出的各模块的有效性，设计了一项消融研究以评估CBM-Net中不同模块的准确率。如表1，包括骨干网络、特征调制模块和梯度调制模块。分别从不同模块组合的效果进行分析，以证明网络整体设计的有效性。为了更好地展示实验效果，CBM-Net采用的主干网络不通过预训练，而是从头开始训练，并对最优结果进行加粗展示。

表1 CBM-Net中两个模块的评估

Table 1 Evaluation of two modules in CBM-Net

骨干网络	特征调制	梯度调制	精度/%	
			RAF-DB	AffectNet-7
√	×	×	86.47	63.13
√	√	×	87.34	64.12
√	×	√	87.57	64.01
√	√	√	**87.71**	**64.33**

从表1可以直观地观察到，在RAF-DB和AffectNet-7数据集上，仅使用特征调制时，精度分别比基础提高了0.87个百分点和0.99个百分点；仅使用梯度调制时，精度分别比基础提高了1.1个百分点和0.88个百分点。这表明了本文提出的各模块的确有效。最终，CBM-Net在没有预训练的情况下，在数据集RAF-DB和AffectNet-7上分别达到了87.71%和64.33%的精度。

3.3.2 评估CBM-Net中各模块类平均精度

为了进一步验证各模块解决类别不平衡问题的能力，本文使用各模块与仅使用ResNet-18作为主干网络的类平均精度对比，以定量地测试各模块对尾部类的性能影响，直观地展示各模块的类别平衡调制效果。其中，在最有代表性的RAF-DB数据集上进行实验，对梯度调制模块中的超参数 a 进行多次不同赋值测试，以挑选最合适超参数 a。最后，将两模块共同使用，以探索CBM-Net的类平衡效果。为了更好地展示对比效果，网络不通过预训练。

如表2，可以直观地观察到，在样本量较少的类Fear、Disgust和Angry(加粗行)，类平均精度均得到了明显的提升，这表明了CBM-Net中各模块在解决类别不平衡问题的有效性。其中，当 a 设置为0.5时，梯度调制效果提升最为显著，为了方便后续实验，当批次大小为128时，本文将 a 统一设置为0.5。最终报告的结果也受 a 为0.5的限制。

3.3.3 评估损失函数的平衡超参数 λ

超参数 λ 控制着分类交叉熵损失函数 L_{CE} 以及特征调制函数 L_{FM} 在训练过程中的占比，为此，依次选取不同的 λ 取值进行实验，以探索 λ 对CBM-Net的影响。同理，网络不通过预训练。

表2 各模块在RAF-DB数据集上的类平均精度

Table 2 Class average accuracy of each module on RAF-DB dataset 单位:%

表情	样本分布	基线值	特征调制	梯度调制					CBM-Net
				a =0.1	a =0.3	a =0.5	a =0.7	a =0.9	a =0.5
Surprise	1 290	89.66	88.45	82.37	85.54	88.45	84.33	91.48	89.22
Fear	**281**	**56.75**	**58.81**	**58.11**	**61.25**	**60.81**	**59.42**	**63.51**	**60.88**
Disgust	**717**	**60.00**	**66.36**	**61.88**	**66.47**	**68.13**	**72.34**	**71.25**	**68.34**
Happy	4 772	94.00	94.52	94.68	94.87	93.84	92.45	90.38	94.68
Sad	1 982	85.56	86.28	86.82	88.87	88.28	86.42	83.05	86.21
Angry	**705**	**78.40**	**80.12**	**79.01**	**79.67**	**79.92**	**79.48**	**79.63**	**80.21**
Neutral	2 524	87.21	87.00	84.85	84.19	85.00	83.21	77.50	87.32
平均值	1 753	78.80	**80.22**	78.25	80.12	**80.63**	79.66	79.55	**80.98**

图3展示了超参数 λ 对CBM-Net的影响,很明显,过大或过小的选取都会降低CBM-Net的性能。当 λ 在0.8到1.2的范围,网络可以获得良好的性能。在后续实验中,本文将 λ 统一设置为1.0。

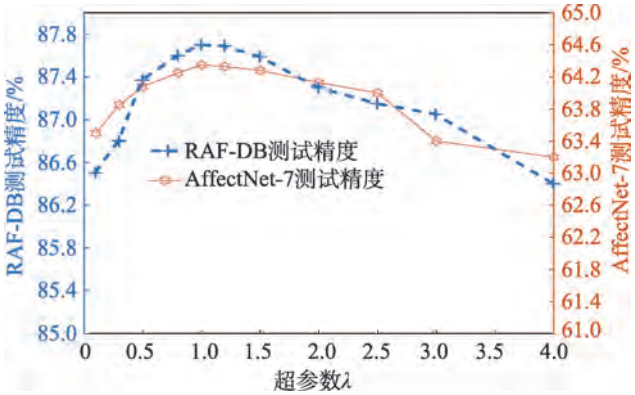

图3 损失函数的平衡超参数 λ 评估

Fig.3 Evaluation on hyperparameter λ of loss function

3.4 特征调制模块的效果可视化

为了直观展示特征调制模块效果,本节采用T-SNE(T-distributed stochastic neighbor embedding)[32]方法对骨干网络提取的图像特征进行可视化。为了比较公平,模型统一采用了在数据集MS-Celeb-1M预训练的ResNet-18作为主干网络,在RAF-DB测试集上进行可视化比较。图4(a)显示了仅使用分类损失 L_{CE} 作为训练约束的特征分布结果。图4(b)显示了除了分类损失 L_{CE} 之外,还添加了特征调制模块的 L_{FM} 作为约束的特征分布结果。

图4(b)展示的带特征调制模块的效果,很明显,最大化类间方向间隔约束使得较小的类也获得更有差异性的特征,不同类在空间分布上保持了平衡。

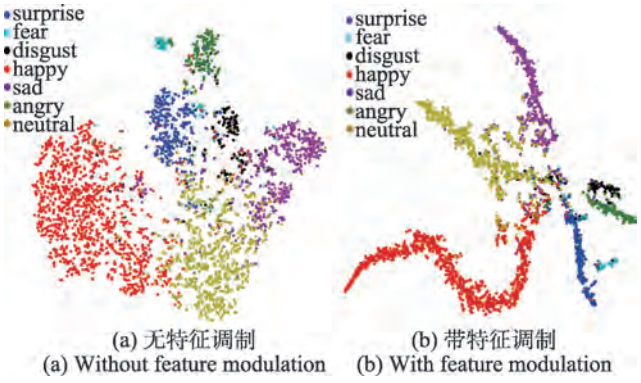

图4 特征调制效果可视化示意图

Fig.4 Diagram of feature modulation effect visualization

3.5 梯度调制模块的效果可视化

为了直观展示梯度调制模块的效果,本节对训练迭代过程中RAF-DB测试集的类平均精度进行了可视化。模型不通过预训练,不使用特征调制模块,直接在ResNet-18的骨干网络下选择是否使用梯度调制模块,以体现本文设计的梯度调制优势。图5(a)显示了仅使用ResNet-18作为模型的类平均精度随epoch的收敛过程。图5(b)显示了除了ResNet-18之外,还添加了梯度调制模块的各类收敛过程。

相较于图5(a)展示的无梯度调制的效果,很明显,图5(b)中带梯度调制的较小类优化更快地得到了收敛。各类的收敛速率更加协调,分类器的优化得到了平衡。此外,较小类的类平均精度得到显著提升,致使总体的平均精度得到提升,这与表2中的结论一致。

3.6 与常规类平衡方法对比

本文设计了一项对比实验,统一使用ResNet-18作为骨干网络,在同样使用本文提出的特征调制损失 L_{FM} 的前提条件下,将本文的梯度调制方法与使用

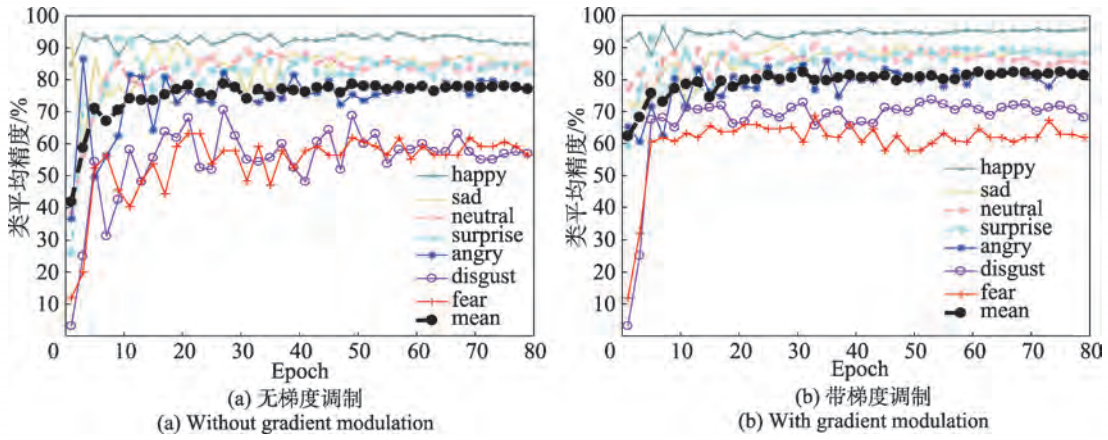

图5 梯度调制效果可视化示意图

Fig.5 Diagram of gradient modulation effect visualization

不同的权重分配方法进行对比，以探索本文提出的梯度调制方法的有效性。如表3，使用权重分配的方法包括常规的带权交叉熵损失函数（W-CE）以及Focal Loss[20]。为了更好地展示对比效果，网络从头进行训练，不通过预训练，并对最优结果进行加粗展示。

表3 CBM-Net与常规类平衡方法的对比

Table 3 Comparison between CBM-Net and conventional class balance methods

方法	准确率/%	
	RAF-DB	AffectNet-7
L_{FM}+W-CE	86.67	62.88
L_{FM}+Focal Loss	86.34	63.52
CBM-Net	**87.71**	**64.33**

从表3可以直观地观察到，在相同实验设置的前提下，本文的梯度调制方法比常规的带权重的交叉熵损失以及Focal Loss效果更好。优异的对比性能得益于本文提出的两阶段的类别平衡调制方法，即分别对特征和分类器进行调制。而直接作用于最终分类损失的类平衡方式影响了模型对图像的表征学习，进而影响了分类器分类性能，这也从侧面印证了本文将图像表征阶段与分类器调制阶段分离的合理性。

3.7 与最先进的方法对比

为了进一步展示CBM-Net的优越性，本节将CBM-Net在RAF-DB、SFEW、CAER-S和AffectNet数据集上与最先进的方法进行定量实验对比，如表4、表5和表6所示。最先进的方法包括SCN（self-cure network）[33]、EfficientFace[34]和DAN（distract your attention network）[37]等方法。SCN通过设计鲁棒损失函数抑制不确定样本参与网络训练。Efficient-Face使用通道注意力以及标签分布学习监督网络训练。DAN的代码可以获得，因此在相同的设置上复现了该算法，但重新实现的结果和报告的结果有一定的差距，这可能是由于不同的数据预处理和运行环境带来的影响。为了公平，本实验采用相同的设置，复现的结果用星号“*”在表中单独标出。CAER-S数据集比较新颖，在此上的研究比较少，因此DAN[37]和Res2Net-50[40]的结果由复现提供。ADDL（adaptive deep disturbance-disentangled learning）[36]在数据集Multi-Pie[43]上

表4 与最先进的方法精度对比

Table 4 Precision comparison with state-of-the-art methods

RAF-DB		SFEW		CAER-S	
方法	准确率/%	方法	准确率/%	方法	准确率/%
SCN[33]	88.14	DAN[37]	50.92*/58.50	CAER-Net[30]	73.51
EfficientFace[34]	88.36	Island loss[38]	52.52	DAN[37]	84.48*
RUL[35]	88.98	Icept-ResVI[39]	51.90	Res2Net-18[31]	85.28
ADDL[36]	89.34	RAN[10]	56.40	Res2Net-50[40]	85.35*
DAN[37]	**89.37***	DMEU[27]	58.34	EfficientFace[34]	85.87
CBM-Net	89.31	CBM-Net	**60.32**	CBM-Net	**86.52**

注：“*”为复现结果，加粗为最优结果。

表5 AffectNet-7数据集上与最先进的方法对比

Table 5 Comparison with state-of-the-art methods on AffectNet-7 dataset

方法	准确率/%
DMEU[27]	63.11
EfficientFace[34]	63.70
KTN[41]	63.97
DAN[37]	**64.83***
CBM-Net	64.62

注:“*”为复现结果,加粗为最优结果。

表6 AffectNet-8数据集上与最先进的方法对比

Table 6 Comparison with state-of-the-art methods on AffectNet-8 dataset

方法	准确率/%
RAN[10]	59.50
EfficientFace[34]	59.89
SCN[33]	60.23
PSR[42]	**60.68**
CBM-Net	60.24

进行了预训练,DAN、Efficient-Face、SCN和DMEU(latent distribution mining and pairwise uncertainty estimation)[27]在MS-Celeb-1M数据集[31]上进行了预训练。CBM-Net和大多数经典算法一样采用的MS-Celeb-1M数据集预训练方式。

表4给出了RAF-DB、SFEW和CAER-S数据集上的定量比较结果。如表4,在经典的RAF-DB数据集上,CBM-Net取得了与最先进的一些方法几乎持平的结果。在SFEW和CAER-S数据集上,CBM-Net都取得了最好的效果,分别为60.32%和86.52%。对比的方法都没有针对类别不平衡问题设计方法,而精度的巨大提高,也进一步说明在自然场景下采集的数据集需要进行类平衡调制。表5和表6给出了AffectNet-7和AffectNet-8数据集上的比较结果,CBM-Net在这些基准测试中同样获得了不错的表现。值得注意的是,RAF-DB与AffectNet均为经典数据集,为了提高识别精度,当前最先进的方法几乎均采用了大网络或设计特殊的注意力方法,如EfficientFace与DAN。而CBM-Net仅采用最基本的ResNet-18模型,在增加有限的计算量和模型参数的基础上,对各类别特征和分类器进行调制,带来了性能大幅度的提升。一方面证明了数据分布不平衡问题确实对模型性能产生了负面的影响,另一方面证明了本文提出的类别平衡方法的有效性。

4 结束语

本文分析了人脸表情识别(FER)中的类别不平衡问题,并提出分别从图像表征阶段与分类器调制阶段来解决这个问题。因此,本文提出了一种类别平衡调制的人脸表情识别方法(CBM-Net),设计了特征调制和梯度调制两个模块。具体的,特征调制模块对图像提取过程进行调制,确保不同类在特征分布上保持平衡。梯度调制模块对分类器梯度进行调制,使得欠优化的小类获得更多的优化。在四个公共数据集上的实验验证了所提出的CBM-Net的有效性和优越性。然而,CBM-Net可能存在的问题是没有脱离权重分配方法范畴,默认了类样本数量等价于类别的信息量。然而,已经有工作指出[21],类别信息的增量随着样本数量的增加而减少,即同类样本存在信息冗余的情况。因此,下一步将着重研究样本数量与类信息量的具体关系,来调制各类别分布。

参考文献:

[1] 洪惠群, 沈贵萍, 黄风华. 表情识别技术综述[J]. 计算机科学与探索, 2022, 16(8): 1764-1778.
HONG H Q, SHEN G P, HUANG F H. Summary of expression recognition technology[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(8): 1764-1778.

[2] 邵志文, 周勇, 马利庄, 等. 基于深度学习的表情动作单元识别综述[J]. 电子学报, 2022, 50(8): 2003-2017.
SHAO Z W, ZHOU Y, MA L Z, et al. Survey of expression action unit recognition based on deep learning[J]. Acta Electronica Sinica, 2022, 50(8): 2003-2017.

[3] DURIC Z, GRAY W D, HEISHMAN R, et al. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction[J]. Proceedings of the IEEE, 2002, 90(7): 1272-1289.

[4] LI B, MEHTA S, ANEJA D, et al. A facial affect analysis system for autism spectrum disorder[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, Taipei, China, Sep 22-25, 2019. Piscataway: IEEE, 2019: 4549-4553.

[5] JEONG M, KO B C. Driver's facial expression recognition in real-time for safe driving[J]. Sensors, 2018, 18(12): 4270.

[6] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, Jun 13-18, 2010. Washington: IEEE Computer Society, 2010: 94-101.

[7] VALSTAR M, PANTIC M. Induced disgust, happiness and surprise: an addition to the mmi facial expression database [C]//Proceedings of the 3rd Workshop on EMOTION, 2010: 65-70.

[8] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with gabor wavelets[C]//Proceedings of the 3rd International Conference on Face & Gesture Recognition, Nara, Apr 14-16, 1998. Washington: IEEE Computer Society, 1998: 200-205.

[9] LI S, DENG W, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2584-2593.

[10] MOLLAHOSSEINI A, HASANI B, MAHOOR M H. AffectNet: a database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.

[11] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests[C]//LNCS 8228: Proceedings of the 20th International Conference on Neural Information Processing, Daegu, Nov 3-7, 2013: 117-124.

[12] DRUMMOND C, HOLTE R C. C4.15, class imbalance, and cost sensitivity: why under-sampling beats over-sampling [C]//Proceedings of the 2003 Workshop on Learning from Imbalanced Datasets II, Washington, Aug 21, 2003: 1-8.

[13] BUDA M, MAKI A, MAZUROWSKI M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks, 2018, 106: 249-259.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[15] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// LNCS 3644: Proceedings of the 2005 International Conference on Intelligent Computing, Hefei, Aug 23-26, 2005: 878-887.

[16] HUANG C, LI Y, LOY C C, et al. Learning deep representation for imbalanced classification[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 5375-5384.

[17] WANG Y X, RAMANAN D, HEBERT M. Learning to model the tail[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2017, Dec 4-9, 2017: 7029-7039.

[18] MALISIEWICZ T, GUPTA A, EFROS A A. Ensemble of exemplar-SVMs for object detection and beyond[C]//Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Nov 6-13, 2011. Washington: IEEE Computer Society, 2011: 89-96.

[19] DONG Q, GONG S, ZHU X. Class rectification hard mining for imbalanced deep learning[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 1869-1878.

[20] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2999-3007.

[21] CUI Y, JIA M, LIN T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 9268-9277.

[22] KANG B, XIE S, ROHRBACH M, et al. Decoupling representation and classifier for long-tailed recognition[C]//Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Apr 26-30, 2020: 1-19.

[23] REN M, ZENG W, YANG B, et al. Learning to reweight examples for robust deep learning[C]//Proceedings of the 35th International Conference on Machine Learning, Stockholm, Jul 10-15, 2018: 4331-4340.

[24] DHALL A, RAMANA MURTHY O V, GOECKE R, et al. Video and image based emotion recognition challenges in the wild: EmotiW 2015[C]//Proceedings of the 2015 ACM International Conference on Multimodal Interaction, Seattle, May 15, 2015. New York: ACM, 2015: 423-426.

[25] LEE J, KIM S, KIM S, et al. Context-aware emotion recognition networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 10143-10152.

[26] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]//Proceedings of the 34th International Conference on Machine Learning, Sydney, Aug 6-11, 2017: 1885-1894.

[27] SHE J, HU Y, SHI H, et al. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. Piscataway: IEEE, 2021: 6248-6257.

[28] 黄浩, 葛洪伟. 强化类间区分的深度残差表情识别网络[J]. 计算机科学与探索, 2022, 16(8): 1842-1849.
HUANG H, GE H W. Deep residual expression recognition network to enhance inter-class discrimination[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(8): 1842-1849.

[29] DENG J, GUO J, VERVERAS E, et al. Retinaface: single-shot multi-level face localisation in the wild[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision

and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 5203-5212.
[30] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.
[31] GUO Y, ZHANG L, HU Y, et al. MS-CELEB-1M: a dataset and benchmark for large-scale face recognition[C]//LNCS 9907: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Cham: Springer, 2016: 87-102.
[32] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.
[33] WANG K, PENG X, YANG J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 6897-6906.
[34] ZHAO Z, LIU Q, ZHOU F. Robust lightweight facial expression recognition network with label distribution training [C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, the 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 3510-3519.
[35] ZHANG Y, WANG C, DENG W. Relative uncertainty learning for facial expression recognition[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 17616-17627.
[36] RUAN D, YAN Y, CHEN S, et al. Deep disturbance-disentangled learning for facial expression recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia, Seattle, Oct 12-16, 2020. New York: ACM, 2020: 2833-2841.
[37] WEN Z, LIN W, WANG T, et al. Distract your attention: multi-head cross attention network for facial expression recognition[J]. arXiv:2109.07270, 2021.
[38] CAI J, MENG Z, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition[C]// Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, May 15-19, 2018. Washington: IEEE Computer Society, 2018: 302-309.
[39] ACHARYA D, HUANG Z, PANI PAUDEL D, et al. Covariance pooling for facial expression recognition[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 367-374.
[40] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652-662.
[41] LI H, WANG N, DING X, et al. Adaptively learning facial expression representation via CF labels and distillation[J]. IEEE Transactions on Image Processing, 2021, 30: 2016-2028.
[42] VO T H, LEE G S, YANG H J, et al. Pyramid with super resolution for in-the-wild facial expression recognition[J]. IEEE Access, 2020, 8: 131988-132001.
[43] GROSS R, MATTHEWS I, COHN J, et al. Multi-PIE[J]. Image and Vision Computing, 2010, 28(5): 807-813.

刘成广(1996—),男,江苏淮安人,硕士研究生,主要研究方向为计算机视觉、情感计算等。
LIU Chengguang, born in 1996, M.S. candidate. His research interests include computer vision, affective computing, etc.

王善敏(1994—),女,江苏扬州人,博士研究生,主要研究方向为计算机视觉、模式识别等。
WANG Shanmin, born in 1994, Ph.D. candidate. Her research interests include computer vision, pattern recognition, etc.

刘青山(1975—),男,江苏南京人,博士,教授,博士生导师,CCF会员,主要研究方向为图像和视觉分析、机器学习等。
LIU Qingshan, born in 1975, Ph.D., professor, Ph.D. supervisor, member of CCF. His research interests include image and vision analysis, machine learning, etc.