arXiv:2302.08650v1 [cs.SD] 17 Feb 2023

高斯平滑的不平衡数据提升语音情感识别

Xuefeng Liang[⋆†] *Hexin Jiang*[⋆] *Wenxin Xu*[⋆] *Ying Zhou*[⋆]

[⋆] 人工智能学院，西安电子科技大学，中国
[†]琶洲实验室，黄埔，中国

摘要

在语音情感识别任务中，模型从数据集中学习情感表示。我们发现，IEMOCAP 数据集中的数据分布极不平衡，这可能会妨碍模型学习更好的表示。为了解决这一问题，我们提出了一种新颖的成对情感数据分布平滑（PDDS）方法。PDDS 认为，现实中的情感数据分布应当是平滑的，因此对情绪对应用高斯平滑，以构建分布更平滑的新训练集。所需的新数据通过 Mixup 数据增强进行补充。由于 PDDS 与模型和模态无关，我们在 IEMOCAP 数据集上使用三种 SOTA 模型对其进行了评估。实验结果表明，这些模型的加权准确率（WA）和非加权准确率（UA）分别提升了 0.2%∼ 4.8% 和 1.5% ∼ 5.9%。此外，消融研究表明，PDDS 的关键优势在于合理的数据分布，而非简单的数据增强。

索引词— 成对情绪数据分布平滑、高斯平滑、Mixup 数据增强、模型与模态无关

1. 引言

语音情感识别（SER）对于理解人类交流具有重要意义。SER 技术已应用于许多领域，例如视频理解 [1],人机交互 [2], 移动服务 [3] 和呼叫中心 [4]。迄今为止，已有大量基于深度学习的情感识别方法被提出 [5, 6]。近年来，多模态情感识别因其表征更丰富、性能更优而受到更多关注 [7, 8, 9] 。这些研究大多为话语分配独热标签。事实上，专家对情感数据的认知往往并不一致，因此独热标签是通过多数投票得出的。图 1(a) 展示了 IEMOCAP 数据集中的情感数据分布 [10]。

后来，一些研究指出，独热标签可能无法很好地表示情绪。它们要么通过多任务学习 [11], 软标签 [12, 13] 和多标签 [14], 来解决这一问题，要么增强模型对

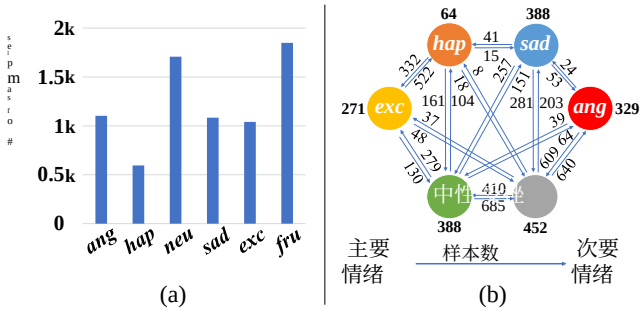
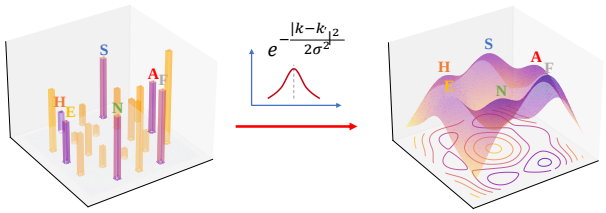


图 1。(a) 标签为独热形式时 IEMOCAP 数据集中的数据分布。(b) IEMOCAP 数据集中清晰样本和模糊样本的统计信息。每个顶点代表一个仅包含清晰样本的情绪类别。粗体数字表示清晰样本的数量。每条有向边表示一组模糊样本，其尾端代表主要情绪，头端代表次要情绪。边上的数字表示模糊样本的数量。

模糊情绪的学习能力，方法是采用动态标签校正 [15]，并通过交互式学习 [16]进行标签重构。这些研究通常将专家投票一致的样本定义为清晰样本，将投票不一致的样本定义为模糊样本。图 1(b) 展示了 IEMOCAP 数据集中清晰样本和模糊样本的统计情况。经统计分析，我们发现数据分布不均衡，尤其是模糊样本。例如，愤怒 与挫败 之间的模糊样本数量很多，而愤怒 与悲伤 之间的模糊样本却很少。同时，快乐 这一清晰情绪类别的样本数量远少于其他清晰情绪类别。我们认为，这种分布并不合理，可能会妨碍模型学习更好的情绪表征。可能的原因是，投票来自少数专家，因此对人群的抽样相当稀疏。

我们认为，现实世界中的清晰与模糊情感数据应遵循平滑的统计分布。基于这一假设，我们提出成对情感数据分布平滑（PDDS），以解决



(a) 平滑前　　(b) 平滑后

图2。平滑前后的数据分布。紫色柱表示清晰样本，橙色柱表示模糊样本。

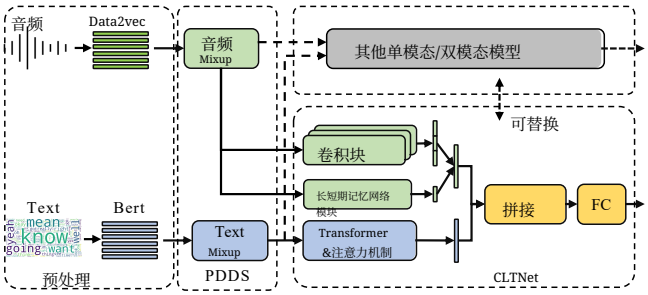


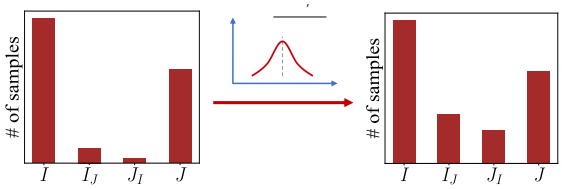
图3。我们方法的框架。它由三个模块组成：（1）数据预处理；（2）PDDS；（3）CLTNet。

数据分布不合理的问题。PDDS 对清晰情感类别对之间的数据分布进行高斯平滑，将模糊样本扩充至合理数量，同时平衡所有类别中清晰样本的数量，使其彼此接近。图 2 展示了平滑前后的数据分布。为补充缺失数据，我们在清晰样本之间进行特征级混合，以扩充数据。

由于 PDDS 与模型和模态无关，我们在三种当前最优方法以及 IEMOCAP 数据集上的单模态和双模态数据上对其进行了评估。结果表明，就加权准确率和非加权准确率而言，这些模型分别提升了 0.2% ∼ 4.8% 和 1.5% ∼ 5.9%。我们提出的 CLTNet 取得了最佳性能。消融研究表明，PDDS 优越性能的本质在于合理的数据分布，而非仅仅增加数据规模。

2. 方法

图 3 展示了我们提出的框架包含三个模块：（1）预处理模块。该模块使用预训练的 data2vec [17] 提取音频特征，并使用预训练的 BERT [18]提取文本特征。（2）成对情绪数据分布平滑模块（PDDS）。该模块平滑不合理的数据分布，是一个可应用于其他情感语音识别方法的插件模块



（a）平滑前　　（b）平滑后

图 4。情绪对数量分布平滑示例。

。（3）CLTNet：一种用于话语级多模态情绪识别的模型。

2.1. 成对数据分布平滑

2.1.1. 数量平滑

假设数据集中有 c 种情绪。由于模糊样本的标签通常来自两种不同的情绪，我们针对每两种情绪 i 和 j 之间四类样本的总体，构建清晰-模糊-清晰分布，其中 i，$j \in \{1, \ldots, c\}$，$i \neq j$。如图 4 所示，它们分别是仅包含情绪 i 的清晰样本I、包含主要情绪 i 和次要情绪 j 的模糊样本 I_J、包含主要情绪 j 和次要情绪 i 的模糊样本J_I，以及仅包含情绪 j 的清晰样本 J。

我们认为，每个情绪对中这四类样本的数量分布在统计上应当平滑，因此将高斯核与该分布进行卷积，得到其平滑版本，

$$n_k = \sum_{k' \in K} e^{-\frac{\|k-k'\|^2}{2\sigma^2}} n_{k'}, \tag{1}$$

其中，$K = \{k-1, k, k+1\}$ 表示$\{I, I_J, J_I\}$或$\{I_J, J_I, J\}$的索引，其中 k是 I_J或 J_I的索引；$n_{k'}$表示平滑前类型 k 中的样本数量，n_k表示平滑后的样本数量。对于清晰样本，如果其在某一情绪类别中的数量过少，则将其扩增，直至达到其他类别中的数量。所有情绪对的平滑后数量分布如图 2(b) 所示。

2.1.2. Mixup 数据增强

平滑数据分布后，原始数据集中某些模糊样本的数量低于预期。为补充数据，我们采用特征级数据增强方法 Mixup [19], 来增强这些样本。Mixup 通过线性组合两个清晰样本及其标签来生成新的训练样本，从而提升模型的泛化能力。其规则如下

$$x_{mix} = p x_\alpha + (1-p) x_\beta, \tag{2}$$

$$y_{mix} = py_\alpha + (1-p)y_\beta,\ p \in [0,1]\,, \tag{3}$$

其中， x_α 和 x_β 分别是主要情绪的一个明确样本和次要情绪的一个明确样本的特征。 x_{mix} 是新样本的特征。 y_α 和y_β 是 x_α 和 x_β的独热标签， y_{mix} 是新样本的标签分布。为避免欠采样，当原始模糊样本的数量达到或超过平滑分布中的数量时，我们使用原始数据。

2.2. CLTNet

为进一步验证 PDDS 的有效性，我们设计了一个简单但有效的话语级多模态融合网络，称为 CLTNet。该网络使用 CNN、LSTM 和 Transformer 提取多模态情绪特征，如图 3 所示。首先，对于音频模态，将声学特征输入三个卷积块，以捕获局部模式。每个卷积块均包含一个一维卷积层和一个全局池化层。为了捕获声学序列中的时间依赖关系，采用了一个长短期记忆层和一个池化层。随后，将得到的 4 个编码特征向量进行拼接，并输入全连接层，以获得音频表示，如下所示：

$$\begin{aligned} h_A &= Concat(x_{conv_A^1}, x_{conv_A^2}, x_{conv_A^3}, x_{lstm_A})W_A + b_A, \\ x_{conv_A^i} &= ConvBlock(x_A), \\ x_{lstm_A} &= LSTMBlock(x_A), \end{aligned} \tag{4}$$

其中， $ConvBlock(\cdot) = MaxPool(Relu(Conv1D(\cdot)))$, $LSTMBlock(\cdot) = MaxPool(LSTM(\cdot))$ $x_A \in \mathbb{R}^{t_A \times d_A}$ 和 $x_{conv_A^i}$， $x_{lstm_A} \in \mathbb{R}^{d_1}$ 分别是卷积块和 LSTM 块的输入特征与输出特征，$W_A \in \mathbb{R}^{4d_1 \times d}$ 和 $b_A \in \mathbb{R}^d$ 是可训练参数。对于文本模态，将文本特征输入由 N 层组成的 Transformer 编码器，以捕获每对文本词语之间的交互。注意力机制 [20] 应用于最后一个块的输出，以聚焦于具有信息量的词语并生成文本表示：

$$\begin{aligned} h_T &= (a_{fuse}^T z_T)W_T + b_T, \\ z_T &= TransformerEncoder(x_T), \\ a_{fuse} &= softmax(z_T W_z + b_z), \end{aligned} \tag{5}$$

其中， $x_T \in \mathbb{R}^{t_T \times d_T}$和 $z_T \in \mathbb{R}^{t_T \times d_2}$分别是 TransformerEncoder 的输入特征和输出特征， $W_z \in \mathbb{R}^{d_2 \times 1}$， $b_z \in \mathbb{R}^1$， $W_T \in \mathbb{R}^{d_2 \times d}$ 和 $b_T \in \mathbb{R}^d$ 是可训练参数。$a_{fuse} \in \mathbb{R}^{t_T \times 1}$ 是注意力权重。最后，将两种模态的表示拼接后，输入三个带有残差连接的全连接层，再经过一个 Softmax层。由于使用标签分布来标注情感数据，我们选择 KL散度损失来优化模型，

$$\begin{aligned} Loss_{KL} &= \sum_{i=1}^{C} y_i log\frac{y_i}{\hat{y_i}}, \\ h &= Concat(h_A, h_L)W_u + b_u, \\ \hat{y} &= Softmax((h + ReLU(hW_h + b_h))W_c + b_c), \end{aligned} \tag{6}$$

其中 $W_u \in R^{2d \times d}$， $b_u \in \mathbb{R}^d$， $W_h \in R^{d \times d}$， $b_h \in \mathbb{R}^d$， $W_c \in \mathbb{R}^{d \times c}$ 和 $b_c \in \mathbb{R}^c$ 为可训练参数，$y \in \mathbb{R}^c$ 为真实标签分布，$\hat{y} \in \mathbb{R}^c$ 为预测标签分布。

3. 实验

3.1. 数据集和评估指标

PDDS 在情感语音识别中最常用的数据集（IEMOCAP）上进行评估[10]。该数据集包含五个会话，其中每个句子均由至少三名专家进行标注。遵循先前的工作 [21], , 我们选择六种情绪：愤怒、快乐、悲伤、中性、兴奋和沮丧。会话 1 至 4 用作训练集，会话 5 用作测试集，且 PDDS 仅应用于训练集。采用加权准确率（ WA，即总体准确率）和非加权准确率（UA，即所有情绪类别上的平均准确率）作为评估指标。

3.2. 实现细节

对于音频，使用预训练的 data2vec [17] 从原始语音信号中提取 512 维特征。帧长和帧移分别设为 25 ms 和 20 ms。对于文本，使用预训练的 BERT [18] 将每个词嵌入为 768 维向量。

在原始数据集中，属于快乐类别的清晰样本数量非常少，因此我们使用 Mixup 增强数据，以平衡所有情绪类别中清晰样本的数量。增强设置为$p = 0.5$, $\alpha = \beta =$ 快乐，数量从 43 增加到 215。随后，使用高斯核 $\sigma = 1$对每种情绪对中的模糊样本应用数据分布平滑。

CLTNet 中三个 CNN 模块的卷积核大小分别为 3、4 和 5，LSTM 中的隐藏单元数为 300。Transformer 编码器有 5 层，每层包含 5 个注意力头。公式（4）和（5）中的嵌入维度d_1， d_2 和 d 分别设为 300、100 和 100。CLTNet 使用 Adam优化器进行训练。学习率、权重衰减和批量大小分别设为 0.001、0.00001 和 64。如果损失连续 10 个训练轮数未下降，训练就会停止；最多训练 30 轮。

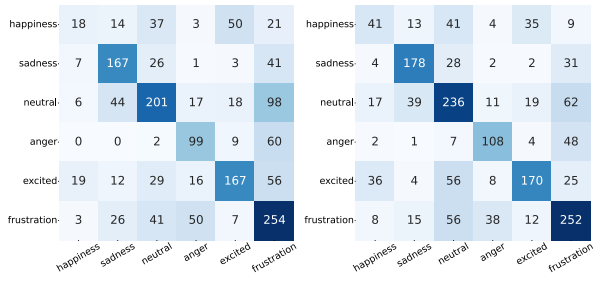
表1。对比使用和不使用 PDDS 的模型。

模型	不使用 PDDS		使用 PDDS	
	WA(%)	非加权准确率(%)	WA(%)	非加权准确率 (%)
SpeechText [7]	55.7	50.6	**58.8**	**55.1**
MAT [8]	54.5	51.0	**58.0**	**55.5**
MCSAN [9]	60.0	56.5	**60.2**	**58.0**
CLTNet（仅音频）	44.9	39.0	**48.4**	**44.9**
CLTNet（仅文本）	50.9	45.8	**51.6**	**50.1**
CLTNet	55.9	52.3	**60.7**	**58.2**

3.3. 验证实验

由于 PDDS 与模型模态无关，我们选用旨在实现话语级情感识别的 SpeechText[7], MAT [8], 和 MCSAN [9], 来评估其有效性。我们提出的 CLTNet 及其单模态分支也接受了测试。实验结果见表1。可以观察到，使用 PDDS 处理训练数据后，所有模型的性能均显著提升，WA 提高了 0.2% ∼ 4.8%，UA 提高了 1.5% ∼ 5.9%。其中，我们提出的 CLTNet 表现最佳。这一结果表明，训练集中的合理数据分布确实有助于模型学习更好的情感表示。

更详细的分析结果，即采用和不采用 PDDS 的 CLT-Net 混淆矩阵，如图5所示。可以看出，大多数情绪类别的分类准确率均有所提高。仅有两个额外样本被误分类为沮丧。然而，混淆矩阵最右侧的列表明，在原始数据集上训练的模型更倾向于将更多样本误分类为沮丧。相比之下，PDDS 显著缓解了这一倾向。



(a) 不使用 PDDS (b) 使用 PDDS

图5。采用和未采用 PDDS 的 CLTNet 混淆矩阵。

3.4. 消融研究

为验证 PDDS 的合理性，我们在 CLTNet 模型上设计并测试了另外两项实验：(a) 仅平衡快乐：仅对

表 2。不同数据增强方法的评估。

Data	加权准确率 (%)	非加权准确率(%)
原始训练数据	55.9	52.3
仅平衡快乐	57.4	54.1
均匀平衡	58.6	56.0
使用 PDDS	**60.7**	**58.2**

快乐 的清晰样本而非模糊样本应用 Mixup 数据增强。(b) 均匀平衡：如果每个类别的清晰样本和每种情绪对的模糊样本数量少于 400，则将其扩增至 400（IEMOCAP 数据集中的大多数样本数量都少于 400）。

结果如表 2 所示。我们可以观察到，三种数据增强方法都提升了模型的性能。与仅平衡快乐 类别相比，同时增强模糊样本和清晰样本有助于提升模型的表现。尽管均匀平衡具有

在规模最大的训练数据集上，PDDS 在 WA 和 UA 上表现最佳，分别提升了 4.8% 和 5.9%。这一结果表明，PDDS 的优势本质上在于合理的数据分布，而非单纯增大数据规模。

我们进一步评估了 Mixup 数据增强中 p 值的有效性，并在图 6 中展示了结果。可以看出，当 $p = 0.8$ 时，效果最佳。

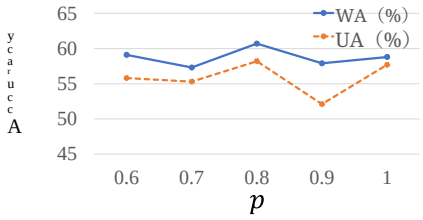


图 6。不同 p 值的影响。

4. 结论

本文针对 IEMOCAP 数据集中的数据分布不平衡问题，提出了成对情感数据分布平滑（PDDS）方法。PDDS 对情绪对应用高斯平滑，构建分布更平滑、更加合理的训练集，并通过 Mixup 数据增强补充所需数据。实验结果表明，PDDS 显著提升了三种当前最优方法的性能。同时，我们提出的 CLTNet 取得了最佳结果。更重要的是，消融研究证实，PDDS 的优势本质上源于合理的数据分布，而非单纯增加数据量。未来，我们将探索更加合理的数据分布，以促进更好的情感表征学习。

5. 参考文献

[1]Xun Gao、Yin Zhao、Jie Zhang 和 Longjun Cai，《电视剧视频中的成对情感关系识别：数据集与基准》，载于 *ACM MM*，2021 年，第 3380–3389 页。

[2] Roddy Cowie、Ellen Douglas-Cowie、Nicolas Tsapatsoulis、George Votsis、Stefanos Kollias、Winfried Fellenz 和 John G Taylor，《人机交互中的情绪识别》，*IEEE* 信号处理杂志，第 18 卷第 1 期，第 32–80 页，2001 年。

[3]Xu Huahu、Gao Jue 和 Yuan Jian，《语音情感识别在智能家用机器人中的应用》，载于 *ICAICI*。IEEE，2010 年，第 1 卷，第 537–541 页。

[4] PurnimaGupta 和 Nitendra Rajput，《呼叫中心监控中的双流情绪识别》，载于 *Proc.Interspeech 2007*。Citeseer，2007 年，第 2241–2244 页。

[5]Jack Parry、Dimitri Palaz、Georgia Clarke、Pauline Lecomte、Rebecca Mead、Michael Berger 和 Gregor Hofer，《跨语料库语音情感识别的深度学习架构分析》，载于 *Proc.Interspeech2019*，2019 年，第 1656–1660 页。

[6] Soujanya Poria、Erik Cambria、Devamanyu Hazarika、Navonil Majumder、Amir Zadeh 和 Louis-Philippe Morency，《用户生成视频中的上下文相关情感分析》，载于 *ACL*，2017 年，第 1 卷，第 873–883 页。

[7] Bagus Tris Atmaja、Kiyoaki Shirai 和Masato Akagi，《利用语音特征和词嵌入进行语音情感识别》，载于 *APSIPAASC*。IEEE，2019 年，第 519–523 页。

[8] Jean-Benoit Delbrouck、No´e Tits 和 St´ephaneDupont，《用于语言-声学情绪识别的 Transformer 调制融合》，载于首届文本之外自然语言处理国际研讨会论文集，2020 年，第 1–10 页。

[9] LicaiSun、BinLiu、Jianhua Tao 和ZhengLian，《用于语音情感识别的多模态交叉与自注意力网络》，载于 *ICASSP*。IEEE，2021 年，第 4275–4279 页。

[10]Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan，“IEMOCAP：交互式情绪双人动作捕捉数据库”，语言资源与评估，第 42 卷，第 4 期，第 335–359 页，2008 年。

[11] Reza Lotfian 和Carlos Busso，“通过多任务学习联合预测类别情绪与主要情绪和次要情绪”，载于《*2018* 年国际语音通信协会年会论文集》，2018 年，第 951–955 页。[12] Haytham M Fayek、Margaret Lech 和 Lawrence Cave- don，“利用深度神经网络对情绪识别中的主观性进行建模：集成模型与软标签”，载于 *IJCNN*。IEEE，2016 年，第 566–570 页。

[13]Atsushi Ando、Satoshi Kobashikawa、Hosana Kamiyama、Ryo Masumura、Yusuke Ijima 和 Yushi Aono，“针对基于深度神经网络的语音情绪分类，使用模糊情绪话语进行软目标训练”，发表于*ICASSP*。IEEE，2018 年，第 4964–4968 页。

[14]Atsushi Ando、Ryo Masumura、Hosana Kamiyama、Satoshi Kobashikawa 和 Yushi Aono，“基于多标签情绪存在模型的语音情绪识别”，发表于 *Interspeech 2019* 论文集，2019 年，第 2818–2822 页。

[15]Takuya Fujioka、Takeshi Homma 和 Kenji Nagamatsu，“考虑情绪标签歧义的语音情感识别元学习”，载于*Proc.Interspeech 2020*，2020 年，第 2332–2336 页。

[16] Ying Zhou、Xuefeng Liang、Yu Gu、Yifei Yin 和 Long- shan Yao，“用于模糊语音情感识别的多分类器交互式学习”，*TASLP*，第 30 卷，第 695–705 页，2022 年。

[17] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli，“data2vec：用于语音、视觉和语言自监督学习的通用框架”，载于 *ICML*，2022 年，第 162 卷，第 1298–1312 页。

[18]Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：用于语言理解的深度双向 Transformer 预训练》，载于*NAACL*，2019 年，第 4171–4186 页。

[19] Hongyi Zhang、Moustapha Cisse、Yann N. Dauphin 和 David Lopez-Paz，《mixup：超越经验风险最小化》，载于 *ICLR*，2018 年。

[20] Zheng Lian、Bin Liu 和 JianhuaTao，《CTNet：用于情绪识别的对话式 Transformer 网络》，*TASLP*，第 29 卷，第 985–1000 页，2021 年。

[21] Devamanyu Hazarika、Soujanya Poria、Rada Mihalcea、Erik Cambria 和 Roger Zimmermann，《ICON：用于多模态情绪检测的交互式对话记忆网络》，载于*EMNLP*，2018 年，第 2594–2604 页。