Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(2), 363-374
Published Online May 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.132038

基于门控复归单位(GRU)和多头注意机制的语音情感识别模型

郭凤婵，吴毅良*，罗序良，刘翠媚

广东电网有限责任公司江门供电局，广东 江门

收稿日期：2024年4月8日；录用日期：2024年5月24日；发布日期：2024年5月31日

摘　要

本研究提出了一种基于门控复归单位(GRU)和多头注意机制的语音情感识别模型。随着人工智能和情感计算的进步，该模型旨在分析语音信号中的情感信息，以识别说话者的情感状态，包括喜怒哀乐等各种情感表达。这一技术在情感智能、智能客服和人机交互等领域有着广阔的应用前景。本研究结合了GRU的时序信息处理能力和多头注意机制对重要特征的关注度提升，构建了一个有效且精确的语音情感识别模型。实验结果表明，此模型在IEMOCAP和Emo-DB数据集上分别实现了81.04%和94.93%的未加权准确率，相较于已有模型有显著提升。此外，该模型还展现出良好的泛化性能和可扩展性，为智能语音交互、情感计算等领域提供了可靠的技术支持。

关键词

语音情感识别(SER)，门控复归单位(GRU)，多头注意机制，Bi-GRU，深度学习

A Speech Emotion Recognition Model Based on Gated Recurrent Units (GRU) and Multi-Head Attention Mechanism

Fengchan Guo, Yiliang Wu*, Xuliang Luo, Cuimei Liu

Guangdong Power Grid Co., Ltd. Jiangmen Power Supply Bureau, Jiangmen Guangdong

Received: Apr. 8th, 2024; accepted: May 24th, 2024; published: May 31st, 2024

*通讯作者。

文章引用: 郭凤婵, 吴毅良, 罗序良, 刘翠媚. 基于门控复归单位(GRU)和多头注意机制的语音情感识别模型[J]. 人工智能与机器人研究, 2024, 13(2): 363-374. DOI: 10.12677/airr.2024.132038

Abstract

This study proposes a speech emotion recognition model based on Gated Recurrent Units (GRU) and a multi-head attention mechanism. With the advancement of artificial intelligence and affective computing, the model aims to analyze emotional information in speech signals to identify the emotional states of speakers, encompassing various expressions such as joy, anger, sadness, and others. This technology holds broad application prospects in affective intelligence, intelligent customer service, and human-computer interaction. Integrating the temporal information processing capability of GRU and the elevated attention to crucial features by the multi-head attention mechanism, an effective and precise speech emotion recognition model is developed. Experimental results demonstrate that this model achieved an unweighted accuracy of 81.04% on the IEMOCAP dataset and 94.93% on the Emo-DB dataset, showing significant improvement compared to existing models. Additionally, the model exhibits good generalization performance and scalability, providing reliable technical support for intelligent speech interaction, affective computing, and related fields.

Keywords

Speech Emotion Recognition (SER), Gated Recurrent Units (GRU), Multi-Head Attention Mechanism, Bi-GRU, Deep Learning

1. 引言

在人工智能和情感计算快速发展的背景下，语音情感识别是指分析语音信号中的情感信息来识别说话者的情感状态，包括喜怒哀乐等多种表达方式。这种技术在情感智能、智能客服和人机交互等领域有着广泛的应用前景。情感分析是自然语言处理(NLP)中的重要任务，旨在让机器理解人类情感。由于语音是人们日常交流的主要方式，其中蕴含着丰富的情感信息。因此，语音情感识别(SER)系统被定义为处理和分类语音信号以侦测内在情感的一系列方法。从语音信号中提取关键的情感信息是语音处理领域备受关注的研究课题。

为了清晰理解语音中的情感变化，提取最相关的声学特征一直是语音情感识别研究中备受关注的课题。随着深度神经网络的应用，情感识别模型有了显著改进。本研究旨在结合门控复归单位(GRU)和多头注意机制，构建一种更加有效、精确的语音情感识别模型。通过引入 GRU，可以充分考虑语音信号中的时序信息，而多头注意机制可以提高模型对重要特征的关注度，进一步提高语音情感识别的准确性和鲁棒性。该研究旨在克服传统语音情感识别模型的局限性，提高情感识别任务的性能，为智能语音交互、情感计算等领域提供更加可靠、高效的技术支持。

GRU 是一种循环神经网络(RNN)的变体，旨在克服普通 RNN 存在的梯度消失和梯度爆炸问题。GRU 通过引入更新门和重置门来控制信息的流动，更好地捕捉时序数据中的长期依赖关系。相较于传统的 RNN 结构，GRU 在语音情感识别等任务中得到了广泛应用，因为它具有更强的记忆和表征能力，并且能够更有效地学习时序数据中的特征。

多头注意机制是一种注意力机制的扩展版本，用于从多个不同的角度对输入数据进行注意力加权的学习和表示。通过使用多个注意力头，模型可以同时关注输入数据的不同部分，并能够对不同特征子空间进行建模，从而提高了模型对输入的表征能力和泛化能力。在语音情感识别任务中，多头注意机制可以帮助模型更好地捕捉语音信号中的关键特征，识别并区分不同的情感表达，从而提高情感识别的准确性和鲁棒性。

综上所述，GRU 和多头注意机制在语音情感识别任务中具有重要意义，能够有效地处理时序数据、提取关键特征，并在模型的训练和推理过程中发挥关键作用。

2. 相关研究

在语音情感识别(SER)中，通常采用传统分类或回归算法以及深度学习方法[1]。传统方法具有较低的算法复杂度和处理速度，主要包括支持向量机(SVM)、高斯混合模型(GMM)、隐藏马尔可夫模型(HMM)和随机森林(RF)。支持向量机旨在找到在样本空间中具有最大间隔的超平面，产生更健壮的分类结果。高斯混合模型通过无监督学习对数据进行分类和转换，为许多后续方法提供借鉴。隐藏马尔可夫模型可以根据观测数据估计和预测未知变量，并提高了模型与观测序列的匹配度。随机森林具有简单的结构和少量计算，可用于分类和回归问题，即使数据集不完整也能保持较高的分类准确性。随着计算设备和传感器水平的提升，越来越多的研究人员选择深度学习方法。相较于传统方法，深度神经网络在训练速度和识别性能方面有更大的优势。

近年来，许多学者在语音情感识别模型上采用了深度学习方法，这有利于更准确地提取语音的情感特征。例如，ED-TTS [2]利用语音情感分析和语音情感识别来区分不同层次的情感，该方法可以获取到帧级的情感信息。Zou 等人[3]提出了一个基于多层次声学信息的端到端语音情感识别系统，利用了 MFCC、频谱图和高级嵌入式声学信息构建多模态特征输入，并使用协同关注机制进行特征融合。Kim 等人[4]利用 CNN 进行新的分类研究，并在句子级分类任务中取得较好的效果。Badshah 等人[5]使用了一个包括矩形滤波器的 CNN 架构，并证明了它在智能医疗中的有效性。另外，在语音情感识别中，递归神经网络(RNN)也被众多学者用于解决情感分析问题。Hasim Sak 等人[6]使用了 LSTM 来处理语音情感识别，并取得了良好的效果。Tao 等人[7]引入了一种 Advanced long short-term memory (A-LSTM)方法，它采用了池化递归神经网络来学习序列，并比简单的 LSTM 具有更好的性能。多头注意力方法被证实在使模型更专注于输出的特定特性方面具有广泛验证[8]。Chung-Cheng Chiu 等人[9]引入了一种多头注意力(MHA)方法来改进自动语音识别(ASR)框架。在文献[10]中，使用多头注意力来提取子空间中不同位置的信息，进一步提高了框架的识别性能。Zhao 等人[11]提出了一种新的 CNN-LSTM 结构，通过学习 Log-Mel 频谱图中的带有时间序列的特征信息，着力解决语音情感识别中的问题。Tara N. Sainath 等人[12]的研究发现，与 DNN 相比，CNN 和 LSTM 在分类应用上对于多种语音识别任务有更好的效果。Chen 等人[13]基于语音和文本信息提出了一种多尺度融合的框架 STSER，并在此基础上引入神经网络(CNN, Bi-LSTM)和注意力机制来对源数据进行训练和分类。

综合而言，当前语音情感识别所面临的问题包括：1) 不一致的特征尺度，对平衡局部和长序列情感特征的重要性具有挑战性；2) 多余的信息影响了语音情感识别的准确性和稳定性；3) 使用过于复杂的模型会在特定任务或数据集上发生过拟合情况。

3. 算法设计

在本节中，我们提出了一种新的语音情感识别模型，该模型结合了双向门控循环单元(Bi-GRU)网络和注意力机制，用于取语音信号的情感特征。如图 1 所示，展示了我们构建的网络结构，包括谱图特征

输入、Bi-GRU 层、多头注意力层和 Softmax 层，用于提取、训练和分类语音情感信息。Bi-GRU 层用于学习长期相关性和上下文特征信息，多头注意力则用于关注与情感相关的特征。最后，Softmax 层用于输出各种情感分类以改善整体性能。

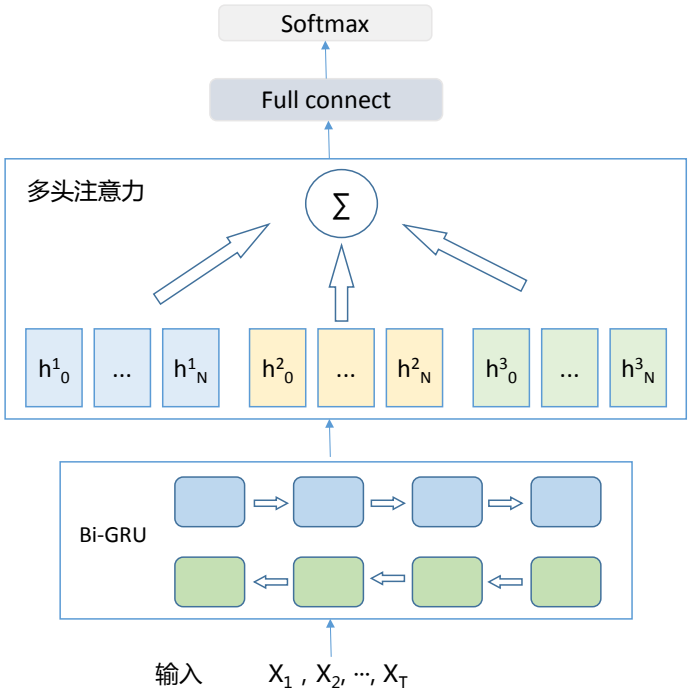

Figure 1. Network Structure based on Bi-GRU and Multi-Head Attention
图 1. 基于 Bi-GRU 和多头注意力的网络结构

3.1. 谱图提取

在研究准备阶段，我们选择了谱图作为提出的语音情感模型的输入信息，因为谱图能够保留丰富的频域信息。为了保留原始语音信号中的情感信息完整性，我们对实验数据集中的语音信号进行了 16 KHz 的采样，并将其组织成单个句子，持续时间从不到一秒到约 20 秒，然后进行短时傅里叶变换获取谱图。实验中，我们使用了 IEMOCAP 和 Emo-DB 语料库，并选择了规模相似的情感子集。这种方法对于全面分析模型对各种情感的分类工作中有较大的性能提升。为更直观地解释实验中语料库中每种情感的比例，我们在表 1 中列出了每种情感所选句子的数量。每个句子至少标记有一种情感类别。对于重叠的加重平均窗口序列，我们把帧步长设置为 10 毫秒，帧长设置为 40 毫秒。在训练中的每个帧中，DFT 的具体计算长度为 1600。换句话说，我们的网格分辨率为 10 Hz。考虑到人声信号频率通常为 300~3400 Hz，并经过实验的重复验证和分析，我们最终选择了 0~4 千赫的频率范围，而忽略了其他范围。在合并短时谱之后，我们获得了一个大小为$(N \times M)$的矩阵；在具体的实验中，变量 N 通常对应于所选时间网格分辨率的语音句子数量，并用于表示选定的时间网格分辨率。变量 M 用于表示实验中的频率网格分辨率。在获得 DFT 数据后，我们将其转换为功率谱，并通过训练数据集的均值和标准差对功率谱进行了归一化处理。

语音样本的句长有所不同，为了提高计算效率，我们按长度进行升序排序。当输入序列具有类似时长的谱图时，我们将它们组织到同一批次中，并在当前批次的谱图中填充 0，直至达到批次中谱图的最大长度。在我们所提出的网络模型训练阶段，我们对一批样本进行统一的并行计算。

Table 1. Detailed Information on the IEMOCAP and Emo-DB Corpora
表 1. IEMOCAP 和 Emo-DB 语料库的详细信息

情感类别	Emo-DB	IEMOCAP
中立	100	100
愤怒	100	100
高兴	100	100
悲伤	100	100
恐惧	100	-
厌恶	100	-
无聊	100	-

3.2. 双向 GRU 层

GRU 网络是 LSTM 的一种改进版本，它的作用是解决长期记忆和梯度消失等问题。相较于 LSTM 网络，它不仅能够获取输入语音的上下文特征信息，而且计算复杂度更低。由于语音信号与时间维度密切相关，在交谈过程中，人们通常基于先前时间的话语信息来思考未来时间内的对话内容。在本文的研究中，GRU 网络充分利用了其在时间维度上的优势。换句话说，实验将首先分析整个话语的情感语调，然后深入分析话语中的局部情感信息特征。实验结果显示，GRU 网络在训练时的算法复杂度更低，并且模型的识别性能也更好。

在传统的 GRU 网络中，通常根据时间规则提取和分析语音中的情感特征，并通过分析先前时刻的情感来确定下一时刻的潜在情感。然而，实践中发现人们日常对话中的情感变化非常复杂，当前语音可能与未来语音相关。例如，人类用一定的情感来表达当前的词语，随着对话的进行，情感会不断变化。考虑到上述情况以及实验中的研究和验证，如果将双向 GRU 层添加到整个网络中，可以更好地弥补单项 GRU 网络造成的部分情感特征损失。在经典的 GRU 网络架构中，主要包括更新门和重置门。直观地看，重置门决定了新的输入信息如何与先前记忆信息相结合，而更新门则定义了保存到当前时间的先前记忆信息的数量。其计算公式如等式(1)和(2)所示。当在时刻 t 输入 x_t 时，双向 GRU 网络可以获取前向和后向信息的隐藏状态 h_t 和 $\tilde{h}_t$。

$$\vec{h}_t = \overrightarrow{GRU}\left(x_t, \vec{h}_{t-1}\right) \tag{1}$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{2}$$

Bi-GRU 的前向和后向传播都包括对重置门 r_t、更新门 z_t 和隐藏状态 $\tilde{h}$ 的计算。它们的计算过程如下：

$$r_t = \sigma\left(W_r \cdot \left[x_t, h_{t-1}\right] + b_r\right) \tag{3}$$

$$z_t = \sigma\left(W_z \cdot \left[x_t, h_{t-1}\right] + b_z\right) \tag{4}$$

$$\tilde{h} = \tanh\left(W \cdot \left[x_t, r_t \otimes h_{t-1}\right] + b\right) \tag{5}$$

其中，σ 表示 sigmoid 函数，W 表示权重，b 表示偏置。时刻 t 的隐藏状态输出为 $\vec{h}_t$ 和 $\overleftarrow{h}_t$，其中 $h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right]$。然后，通过组合 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 来获取全局上下文信息，以捕获语音上下文特征信息向量。

3.3. 多头注意力机制

近年来，深度学习网络的注意力机制在语音情感识别领域得到了广泛应用，以确保分类器能根据输

入的每个部分的注意力权重关注于给定样本的具体位置。注意力机制主要包含查询(query)、键(key)、值(value)以及输出(output)这些向量。一般来说，典型的注意力机制通过计算键和值的映射来自动学习并计算输入数据对输出数据的影响。在训练过程中，模型根据输入数据和神经元的历史输出来预测当前时间步进，最终得到输入数据每个维度的权重。这一计算过程中，每个值都需要利用查询和匹配键的兼容性函数来计算。传统的注意力机制采用的方法是将获得的上下文向量集中在输入序列的特定表示子空间上。然而，以这种方式获得的上下文向量仅能反映出输入语义的一个方面。一般来说，一个句子可能涉及多个语义空间，特别是对于较长的输入序列。

为了在模型训练过程中实现更好的分类性能，我们的双向 GRU 模型引入了多头注意力机制。多头注意力模型可以根据头数表示不同的子空间，并在不同位置获得公共的注意力信息。注意力机制算法的具体计算方法如下公式(6)~(7)所示。

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\cdots,head_h\right)W^O \tag{6}$$

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{7}$$

其中 $W_i^Q\in R^{d\times d_q}$， $W_i^K\in R^{d\times d_k}$， $W_i^V\in R^{d\times d_v}$， $W_i^O\in R^{hd_v\times hd_1}$。

在获得 GRU 网络的输出向量后，我们发现对于 d_q， d_k 和 d_v 维度进行 h 次线性投影是有益的。这种线性投影分别通过不同的且可学习的查询(query)、键(key)、和值(value)的向量计算完成的。然后，我们通过并行地对每个查询(query)、键(key)和值(value)的投影执行注意力函数，生成 d_v 维度的输出值。最后，这些向量经过层级连接和再次投影以获得最终的数值。

模型的整体流程如算法 1 所示。

算法 1. 模型算法的伪代码。

输入：时间－频率特征量 $X_1,X_2,\cdots,X_T$

输出：情感类别及概率 Y,P

// Bi-GRU 算法(forward and backward)

$\vec{h}_t\leftarrow\overrightarrow{GRU}\left(x_t,h_{t-1}\right)$

$\overleftarrow{h}_t\rightarrow\overleftarrow{GRU}\left(x_t,\tilde{h}_{t-1}\right)$

$h_t=\left[\vec{h}_t,\overleftarrow{h}_{t-1}\right]$

// 多头注意力算法

$Q=h_t\cdot W_q,K=h_t\cdot W_k,V=h_t\cdot W_v$

$$AttentionScore_i=softmax\left(\frac{QW_{Q_i}^{\mathrm{T}}K^{\mathrm{T}}}{\sqrt{d_k}}\right)$$

$head_i=AttentionScore_i\cdot V$

$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\cdots,head_h\right)W^O$

// 线性层和输出层

$Y'=fullconnect\left(MultiHead\left(Q,K,V\right)\right)$

$Y=softmax\left(Y'\right)$

4. 实验与结果分析

如前文所述，我们选择了 IEMOCAP 和 Emo-DB 数据集来评估提出的系统。同时，为了验证模型对

不同任务的可扩展性，我们还将该模型应用于语音情感分析任务。我们选取了用于语音情感分析任务的数据集 CH-SIMS [14]和 MOSI [15]。

4.1. 实验设置

我们选择了 IEMOCAP 数据集[16]作为本研究的一个数据来源，该数据集主要用于研究多模态表达中的二元交互。该语料库包括 10 名受试者的面部动作捕捉以及相关的音频/视频录音，总时长达 12 小时。每个会话涉及不同的两人组，演讲者们在脚本和情感情境提示词的引导下进行自发即兴对话中表现出情感。

另一个用于本次实验的语音信号样本输入是 EMO-DB 数据集[17]。我们选择这个数据集的主要原因是其语音特征的多样性、情感类型的丰富性，以及在语音情感识别领域的广泛应用。该数据集包含 535 个德语音频片段，涵盖了愤怒、悲伤、恐惧、中立、高兴、厌恶和无聊等七种情感类别。

在实验中，我们采用了排除一人(leave-one-speaker-out, LOSO)的方法。即在每次模型训练与测试的过程中，选择一个人的全部语音作为测试集，其他人的语音作为训练集。这种方法能更准确地验证算法是否能学习情感特征，而且这些特征与发言者无关。为了提高实验结果中情感分类的准确性，并减少意外事件引起的偏差，我们采用了交叉验证方法，选择 N 个人的语音样本作为测试集，并计算结果的统计平均值作为一个完整的实验结果。

在本研究中，我们使用了 NVIDIA GeForce RTX 4090 24G GPU 平台和 python3.9 + pytorch1.12.1 + cuda11.6 深度学习框架来搭建模型训练环境。我们采用了 Adam 优化器来优化模型的参数，损失函数采用了 L1 损失，激活函数则采用了 ReLU 函数。鉴于 IEMOCAP 和 Emo-DB 数据集之间的差异，我们分别设置了两组超参数，具体设置如表 2 所示。我们遵循了研究人员常用的数值来设置超参数。针对 IEMOCAP 数据集，我们采用了较低的学习率和更多的注意力头以及较多的训练周期，因为 IEMOCAP 数据集的数据量较大。同时，为了减少过拟合现象，我们将 dropout 比率增加到了 0.3。

Table 2. Hyperparameter Settings for the Model
表 2. 模型的超参数设置

超参数	IEMOCAP	Emo-DB
Learning rate	1×10^{-5}	1×10^{-4}
Batch size	32	32
Attention dropout	0.2	0.2
Head_num	16	8
Dropout (Output)	0.3	0.2
Epochs	100	80

4.2. 实验结果

我们使用测试集中所有样本未加权准确度(Unweighted Accurate, UA)来评估了五组情感识别实验的结果。作为用于建模情感特征的分类器，多头注意力层之后连接了两个全连接层和一个 softmax 层。其余部分采用了 Bi-GRU 与多头注意力结合的模型。在实验中，我们设置了 4、8、16 和 32 个注意力头，旨在探究注意力头数量对多头注意力情感识别的影响。另外，我们将只使用表 3 中的 Bi-GRU 网络模型作为实验基准。

Table 3. Comparison of UA results on IEMOCAP and Emo-DB datasets
表 3. IEMOCAP 和 Emo-DB 数据集上的 UA 结果比较

模型	IEMOCAP	Emo-DB
Bi-GRU	77.01%	89.22%
GRU + Head_4	80.47%	90.66%
GRU + Head_8	89.89%	94.93%*
GRU + Head_16	81.04%*	92.81%
GRU + Head_32	79.32%	91.03%

在 Emo-DB 和 IEMOCAP 语料库中，我们观察到 Bi-GRU 模型的预测值与提出的模型之间存在显著差异。从表 3 中可以看出，我们在模型训练期间获得的未加权准确度(UA)在不同的语料库中并不完全相同。具体来说，在独立于说话者的音频分类任务中，当注意力头的数量为 16 和 8 时，最佳性能分别达到了 81.04%和 94.93%。

为了全面分析整个实验过程，我们提供了我们的网络(head = 16)在 IEMOCAP 语料库上的混淆矩阵，详见图 2，并提供了我们的网络(head = 8)在 Emo-DB 语料库上的混淆矩阵，详见图 3。针对独立于说话者的 IEMOCAP 和 Emo-DB 数据集，新模型对于中立和悲伤情感具有良好的识别率。由于语料库中有更多的中立情感样本，训练效果更好。在频谱图上，悲伤情感具有更明显的特征，因此模型对其有较好的识别能力。此外，在 IEMOCAP 数据集中，快乐和愤怒样本在情感特征上相似度较高，因此识别效果较低。而在 Emo-DB 数据集上，恐惧、悲伤和无聊情感获得了较高的识别准确率。

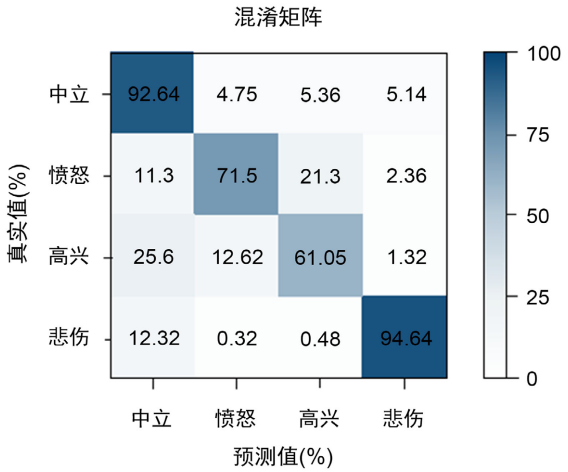

Figure 2. The confusion matrix of the model (head_num = 16) on the IEMOCAP corpus in this paper
图 2. 本文模型(head_num = 16)对 IEMOCAP 语料库的混淆矩阵

为了全面展示模型的整体性能，我们在 IEMOCAP 和 EMO-DB 这两个数据集上计算了总体准确度、精确度和召回率，具体数据如表 4 所示。此外，不同情感的识别准确度详细信息也列在了表 5 中。

与此同时，我们将本文的模型与常用的深度学习模型进行比较，实验结果如表 6 所示。

为了研究所提出的模型的泛化性能和可扩展性，我们还对 CH-SIMS 和 MOSI 数据集进行了语音情感分析实验。具体的实验结果见表 7。我们评估了模型在语音情感分析任务中的性能，考察指标包括二元

准确度(Acc-2)、五分类准确度(Acc-5)、F1 分数、平均绝对误差(MAE)和皮尔逊相关系数(Corr)。除 MAE 外，指标数值越高表示性能越好。

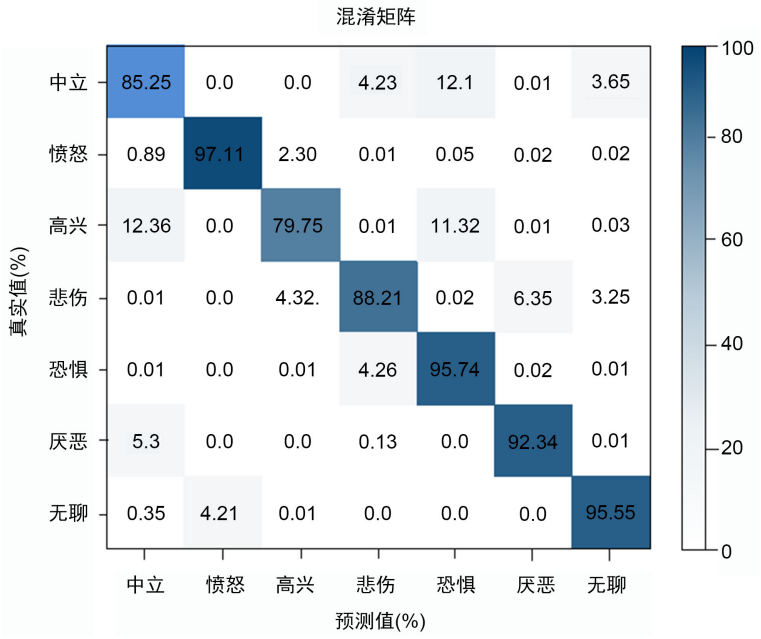

Figure 3. Confusion Matrix of the Model (head_num = 8) on the Emo-DB Corpus in This Paper
图 3. 本文模型(head_num = 8)对 Emo-DB 语料库的混淆矩阵

Table 4. Overall results on IEMOCAP and Emo-DB datasets
表 4. IEMOCAP 和 Emo-DB 数据集上的总体结果

结果	IEMOCAP	Emo-DB
精确度	89.33%	80.57%
准确度	81.04%	94.93%
召回率	76.20%	86.20%

Table 5. Recognition accuracy of different emotion categories on IEMOCAP and Emo-DB datasets
表 5. IEMOCAP 和 Emo-DB 数据集上不同情感类别的识别准确度

情感类别	IEMOCAP	Emo-DB
中立	85.25%	92.64%
愤怒	97.11%	71.50%
高兴	79.75%	61.05%
悲伤	88.21%	94.64%
恐惧	95.74%	-
厌恶	92.34%	-
无聊	95.55%	-

Table 6. Accuracy of different models on the IEMOCAP and Emo-DB datasets
表 6. IEMOCAP 和 Emo-DB 数据集上不同模型的识别准确度

算法	IEMOCAP	Emo-DB
CNN	75.80%	76.58%
LSTM	76.10%	83.73%
RNN	76.22%	85.91%
CNN + LSTM	78.17%	89.79%
Bi-GRU	77.01%	89.22%
Self-Attention	77.87%	88.91%
Bi-GRU + Self-Attention	79.21%	90.58%
Multi-head Attention	78.13%	89.37%
本文算法	81.04%	94.93%

Table 7. Experimental results of the model in speech emotion analysis task
表 7. 模型在语音情感分析任务中的实验结果

评估指标	CH-SIMS	MOSI
Acc-2	88.43%	86.29%
F1-score	87.97	86.31
MAE	0.642	0.968
Corr	0.578	0.657
Acc-5	76.19%	72.12%

4.3. 结果分析

根据表 4，本文算法模型在两个数据集上表现出的性能相对良好。从结果来看，IEMOCAP 和 Emo-DB 对负面情感的识别准确度高于正面情感的识别准确度。我们分析认为，这可能是因为负面情感在语音中具有更明显的特征，而正面情感更容易被识别为中立情感。

根据表 6，我们本文的方法与常用的深度学习模型相比具有更高的准确度。这是由于模型中 Bi-GRU 和多头注意力的共同作用，使模型能够关注语音序列的整体和局部特征。该模型结合了循环神经网络和多注意力的优势。与 Bi-GRU、自注意力和多头注意力等独立模型相比，我们的模型补充并融合信息，从而使语音情感识别结果更加稳定。

根据表 7，该模型在语音情感分析任务中也能取得相对良好的结果。这是因为模型中的 Bi-GRU 层和多头注意力层充分提取了语音信号中的信息，并采用合理的丢弃概率增加了随机性。

根据使用方差分析(ANOVA)统计检验，基于 Bi-GRU 和多头注意力的模型相较于单独使用 Bi-GRU 网络在 Emo-DB 数据集($F(1, 698) = 10.0243$, $p < 0.05$)和 IEMOCAP 数据集($F(1, 198) = 8.4174$, $p < 0.05$)上均显示出显著改进。

通过观察图 4，我们可以看到随着注意力头数的变化，语音情感识别模型的准确度呈现先上升后下降的趋势。这是因为随着注意力头数的增加，模型的复杂度也相应增加。这使得模型更加详细地关注情感信息，但也可能导致模型在学习权重时产生偏差。

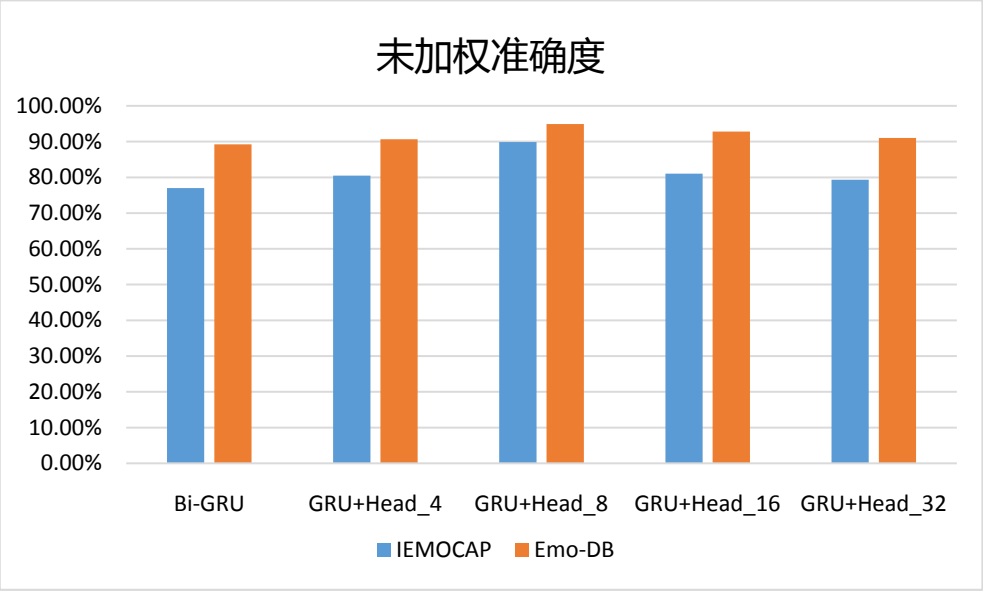

Figure 4. Histogram of Unweighted Accuracy (UA) on IEMOCAP and Emo-DB datasets
图 4. IEMOCAP 和 Emo-DB 数据集上的未加权准确度(UA)直方图

5. 结论

本文提出了一种将 GRU 网络和多头注意力机制相结合的语音情感识别模型。模型的未加权准确率在 IEMOCAP 和 Emo-DB 数据集上分别达到了 81.04%和 94.93%。相较于文献[18]的结果，本文算法在 IEMOCAP 数据集上有 10.82%的提升；同时，与 Mustaqeem 等人[19]的研究结果相比，我们在 IEMOCAP 和 Emo-DB 数据集上分别提高了 2.79%和 3.36%。此外，与原始的模型(Bi-GRU)相比，我们的模型分别提高了 4.40%和 5.71%。与目前大多数先进的语音情感识别模型相比，我们的模型在 IEMOCAP 和 Emo-DB 数据集上具有更好的识别率。通过与常见的深度学习方法进行比较，基于 Bi-GRU 和多头注意力方法的情感识别模型表现出最佳的识别效果。此外，我们还将该模型应用于语音情感分析任务，并在不同数据集上实现了更稳定的预测结果。详细的评估指标被使用以对模型进行更全面的分析，以防止模型在特定参数上过拟合。

尽管结果显示出模型对特定情感的识别准确性并未得到显著提高，我们推测这可能是由于中立情感的语音特征不太明显，且我们的模型在特征融合方面并未采用复杂的方法。未来的工作将会对特征融合方面的进一步研究。此外，多模态情感分析备受关注，因为它具有更丰富的情感信息，对预测和识别人类情感更为全面。因此，利用多任务学习方法、迁移学习或联合学习等先进方法，我们的模型可以得到改进并应用于更多应用场景。

基金项目

本文由“南网高层次人才特殊支持计划”项目资助。

参考文献

[1] 耿磊, 傅洪亮, 陶华伟, 等. 基于动态卷积递归神经网络的语音情感识别[J]. 计算机工程, 2023, 49(4): 125-130. https://doi.org/10.19678/j.issn.1000-3428.0064054

[2] Tang, H., Zhang, X., Cheng, N., Xiao, J., Wang, J. (2024) ED-TTS: Multi-Scale Emotion Modeling Using Cross-Domain Emotion Diarization for Emotional Speech Synthesis. Seoul, 14-19 April 2024, 12146-12150.

https://doi.org/10.1109/ICASSP48485.2024.10446467

[3] Zou, H., Si, Y., Chen, C., *et al.* (2022) Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information. *ICASSP* 2022-2022 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 7367-7371. https://doi.org/10.1109/ICASSP43922.2022.9747095

[4] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing*, Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[5] Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S. and Baik, S.W. (2017) Deep Features-Based Speech Emotion Recognition for Smart Affective Services. *Multimedia Tools and Applications*, **78**, 5571-5589. https://doi.org/10.1007/s11042-017-5292-7

[6] Sak, H., Senior, A., Rao, K., *et al.* (2015) Learning Acoustic Frame Labeling for Speech Recognition with Recurrent Neural Networks. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), South Brisbane, 19-24 April 2015, 4280-4284. https://doi.org/10.1109/ICASSP.2015.7178778

[7] Tao, F. and Liu, G. (2018) Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition. *Proceedings of the* 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, AB, 15-20 April 2018, 2906-2910. https://doi.org/10.1109/ICASSP.2018.8461750

[8] Moritz, N., Hori, T. and Roux, J.L. (2019) Triggered Attention for End-to-end Speech Recognition. *Proceedings of the ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, UK, 12-17 May 2019, 5666-5670. https://doi.org/10.1109/ICASSP.2019.8683510

[9] Chiu, C.C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., *et al.* (2018) State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. *Proceedings of the* 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, AB, 15-20 April 2018, 4774-4778. https://doi.org/10.1109/ICASSP.2018.8462105

[10] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the Neural Information Processing Systems*, Long Beach, CA, 4-9 December 2017, 1-11.

[11] Zhao, J., Mao, X. and Chen, L. (2019) Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks. *Biomedical Signal Processing and Control*, **47**, 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

[12] Sainath, T.N., Vinyals, O., Senior, A. and Sak, H. (2015) Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks. *Proceedings of the* 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), South Brisbane, 19-24 April 2015, 4580-4584. https://doi.org/10.1109/ICASSP.2015.7178838

[13] Chen, M. and Zhao, X. (2020) A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. *Proceedings of the Interspeech* 2020, Shanghai, 25-29 October 2020, 374-378. https://doi.org/10.21437/Interspeech.2020-3156

[14] Yu, W., Xu, H., Meng, F., *et al.* (2020) Ch-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-Grained Annotation of Modality. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistic*, July 2020, 3718-3727. https://doi.org/10.18653/v1/2020.acl-main.343

[15] Zadeh, A., Zellers, R., Pincus, E., *et al.* (2016) Mosi: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv:1606.06259.

[16] Busso, C., Bulut, M., Lee, C.C., *et al.* (2008) IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, **42**, 335-359. https://doi.org/10.1007/s10579-008-9076-6

[17] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J.G. (2001) Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, **18**, 32-80. https://doi.org/10.1109/79.911197

[18] Latif, S., Rana, R., Khalifa, S., Jurdak, R. and Schuller, B. (2022) Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition. IEEE Trans. *Affective Computing*, **14**, 1912-1926. https://doi.org/10.1109/TAFFC.2022.3167013

[19] Mustaqeem, Sajjad, M. and Kwon, S. (2020) Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*, **8**, 79861-79875. https://doi.org/10.1109/ACCESS.2020.2990405