一项实证研　　究与改进：语音情感识别

Zhen Wu、*Yizhe Lu*、*Xinyu Dai*

南京大学新型软件技术国家重点实验室；新型软件技术与产业化协同创新中心，
南京 wuz@nju.edu.cn、luyz@smail.nju.edu.cn、daixinyu@nju.edu.cn

arXiv:2304.03899v1 [cs.CL] 8 Apr 2023

摘要

多模态语音情感识别旨在从音频和文本中识别说话者的情绪。以往的研究主要着眼于利用先进网络对不同模态的信息进行建模与融合，以提升性能，却忽视了不同融合策略对情感识别的影响。在本研究中，我们关注一个简单却重要的问题：如何融合音频和文本模态信息，才能更有助于这一多模态任务。此外，我们提出了一种通过透视损失改进的多模态情感识别模型。实证结果表明，我们的方法在 IEMOCAP数据集上取得了新的最先进结果。深入分析解释了改进后的模型为何能够提升性能并优于基线模型。

索引词— 语音情感识别，多模态融合，视角损失，深度学习

1. 引言

语音情感识别是情感计算中的一项重要子任务，旨在从用户的话语中识别说话者的情绪 [1]。近年来，由于这项任务在社交媒体分析、人机交互和客户服务等领域具有广泛应用，因此越来越受到研究人员和业界的关注。

早期研究使用纯音频信息来识别语音情绪 [2, 3, 4]。例如，Mirsamad 等人 [2]采用基于注意力机制的 RNN 来捕捉情绪感知特征。Gao 等人 [4] 通过对抗神经网络消除音频中不同说话人的影响。事实上，语音天然包含多模态的音频和文本信息，分别从语音学和语言学角度体现情绪。认识到这一特点后，后续研究者使用多模态信息构建神经模型。这些研究 [5, 6, 7, 8, 9] 广泛采用交叉注意力融合来自不同模态的信息，在准确率上显著优于单模态方法。

Zhen Wu 和 Yizhe Lu 对本文贡献相同。本研究得到中国国家自然科学基金（项目编号：62206126 和 61976114）资助。我们的代码可在 https://github.com/Luyizhe/SpeechEmotion 获取。

上述研究均聚焦于单个话语的情感识别。考虑到话语之间的关联，近期研究开始转向更贴近实际场景的对话多模态情感识别。Poria 等人 [10] 采用分层建模，先提取话语级特征，再聚合对话级信息。在后续研究中 [1], ，他们提出一种基于注意力的融合（AT 融合）方法 [1]，以获取不同模态的话语表示，并通过额外的 LSTM [11] 和自注意力机制 [12]获得全局信息。此外，图神经网络也被用于融合音频和文本模态特征 [13, 14]。

尽管现有方法有效，但它们依赖先进的神经网络来取得显著成果，却忽视了不同模态的融合方式对情感识别的影响。在本研究中，我们重新审视这一问题，并探讨一个关键问题：哪种融合方式更适用于多模态情感识别模型？

为实现这一目标，我们从现有研究中提取出一个通用的单模态音频/文本框架作为基本组件，并研究在不同层融合多模态信息的各种方式所产生的影响。我们发现，一些简单操作已取得极具竞争力的性能。此外，我们提出了一种结合视角损失的改进多模态模型，并在 IEMOCAP [15] 数据集上取得了新的最先进结果。实证结果和分析解释了多模态模型及其改进版本能够取得更佳性能的原因。

2. 方法

我们首先定义对话中的情感识别任务，然后展示不同的融合方法。最后，我们介绍视角损失。

2.1. 问题定义

一组对话 $\mathcal{U} = \{U_1, U_2, ..., U_N\}$ 包含 N段对话。每段对话 $U = \{(u_1, e_1), (u_2, e_2) ....., (u_S, e_S)\}$由 S 个话语组成，而 u_i 是其第 i_{th} 个话语。话语 u_i 的情绪标签是 $e_i \in${开心、悲伤、中性、愤怒}。

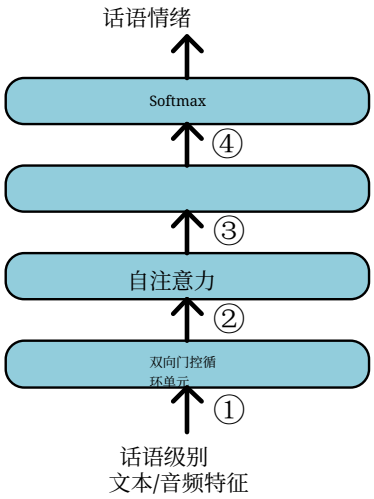

图 1。单模态结构概览。

任务是构建一个模型来拟合 N 带标签的对话 $\mathcal{U}$，然后预测测试集中的对话情绪标签。

2.2. 单模态框架

在应用不同的融合方法之前，我们先介绍单模态音频/文本框架。

如图 1 所示，通用单模态框架是从现有的多模态框架 [1, 10, 16] 中提取的。首先，我们提取话语级音频特征或文本特征（见第 3.2 节）。对于单模态特征，我们随后应用双向 GRU（双向门控循环单元）对所有话语进行编码，并捕获对话中的过去和未来信息。接下来，我们将双向门控循环单元层的输出作为自注意力模块 [1], 的输入，该模块从相关话语中获取丰富的上下文信息。之后，自注意力模块的输出被送入全连接（FC）层，该层将高层特征映射到情绪类别的维度。**请注意**，FC 层的输出也称为 logits。最后，可以应用 softmax 层将 logits 转换为概率，用于单模态情感分类。

2.3. 融合方法

单模态框架包含多个堆叠组件。为分析不同层中多模态融合的效果，我们考察了四个位置： ① 早期融合（EF）、 ② 中间融合（MF）、 ③ 后期融合（LF）和 ④ Logits 融合（LGF），如图 1 所示。

在融合策略方面，文献中多模态融合通常有三种方法，分别是相加融合、拼接融合和 AT 融合。为了清晰地描述这些方法，我们以 Logits 融合（LGF）为例，介绍一些必要的符号。具体而言， $a_i \in \mathbb{R}^d$ 和 $t_i \in \mathbb{R}^d$ 分别表示融合前单模态位置 ④ 处的音频表示和文本表示。这里，音频表示和文本表示的维度相同，均为 d，因为在提取多模态特征时（第 3.2 节），我们会较早地将原始音频和文本特征映射到相同维度。

ADD-Fusion. 通过这种方式，我们直接将两个不同模态的特征相加，以获得融合表示 $z_i \in \mathbb{R}^d$。过程如下：

$$z_i = a_i + t_i. \tag{1}$$

拼接融合. 在此策略中，将不同模态的特征拼接在一起，然后映射到目标维度。融合后的表示 $z_i \in \mathbb{R}^d$ 如下：

$$z_i = W_f \cdot [a_i; t_i], \tag{2}$$

其中， $W_f \in \mathbb{R}^{d \times 2d}$ 和 $[;\ ]$ 表示拼接。

AT 融合. AT 融合以加权求和的形式融合不同模态的特征，其中权重衡量模态的重要性。具体计算过程与 [1]的融合过程相同。

2.4. 视角损失

通常，融合表示 z_i 可直接用于情感分类。相应的交叉熵损失记为**单一损失**：

$$\mathcal{L}_s = \sum_{U \in \mathcal{U}} \sum_{i=1}^{N} \sum_{}^{S} -\log P(\hat{y}_i = y_i), \tag{3}$$

其中，$\hat{}y_i$ 和 y_i 分别是对话 U中话语 u_i 的预测标签和真实标签。

众所周知，音频和文本模态包含多种不同特征，例如语音学和语言学。受相关工作 [17], 的启发，我们分别为音频和文本模态引入了额外的情感分类交叉熵损失，以鼓励模型在融合多模态信息时保留有助于情绪识别的单模态特征。最后，我们将这两个单模态损失与上述**单一损失**结合，构成**视角损失**：

$$\mathcal{L}_a = \sum_{U \in \mathcal{U}} \sum_{i=1}^{N} \sum_{}^{S} -\log P(\hat{y}_i^a = y_i), \tag{4}$$

$$\mathcal{L}_t = \sum_{U \in \mathcal{U}} \sum_{i=1}^{N} \sum_{}^{S} -\log P(\hat{y}_i^t = y_i), \tag{5}$$

$$\mathcal{L}_p = \mathcal{L}_a + \mathcal{L}_t + \mathcal{L}_s, \tag{6}$$

其中，$\hat{}y_i^a$和$\hat{}y_i^t$ 分别表示音频模态和文本模态的预测标签。

模态	WA(%)
音频	71.19
Text	82.58

表 1. 第 2.2 节中提到的音频单模态框架和文本单模态框架在语音情感分类任务上的 WA(%) 结果。

3. 实验设置

3.1. 数据集

我们在 IEMOCAP [15] 数据集上进行实验。该数据集由 10 位说话人在 5 个会话中采集，包含多个小时、涵盖各种对话场景的数据。按照以往研究 [10, 18], 的做法，我们采用四个类别，将开心 和兴奋 合并为开心 类别。因此，最终数据集包含 5531 条话语（开心：1636，中性：1084，愤怒：1103，悲伤：1708）。实验中，我们将 $1-4$ 个会话设为训练集，第 5 个会话设为测试集。

3.2. 特征提取

在本研究中，我们使用不同的工具提取单模态音频和文本话语特征。

3.2.1. 音频特征

按照文献 [19], 中的方法，我们使用带有配置文件 *IS13ComParE* 的软件 openSMILE，为每个话语提取 6373 个特征。随后对特征进行 Z 标准化，以实现归一化。然后，我们使用一个全连接层将音频向量的维度降至 100，并将其作为音频最终的话语级特征。

3.2.2. 文本特征

对于每个话语，我们使用预训练的 BERT$_{base}$ 对文本进行编码，并对所有词元表示进行均值池化，以获得一个 768 维向量。之后，我们使用一个全连接层将该向量降至 100 维，作为文本模态的话语特征。

3.3. 超参数和指标

我们将双向门控循环单元网络的隐藏状态设为100维。自注意力层包含100维状态和4个注意力头。我们使用ReLU作为非线性激活层。采用Adam优化来优化模型，学习率设为0.001，衰减设为0.00001。模型训练150轮，批大小为20，丢弃法的参数为 $p = 0.2$，并采用权重为0.00001的 $L2$ 正则化。为获得统计可信结果，我们使用不同的初始

FL	FM	WA(%) $\mathcal{L}_s$	WA(%) $\mathcal{L}_p$
EF	ADD-Fusion	83.95±0.64	84.57±0.52
	拼接融合	83.12±0.85	83.68±0.52
	AT 融合	83.12±0.59	83.29±0.51
MF	ADD-Fusion	83.86±0.54	85.23±0.54
	拼接融合	83.25±0.79	84.89±0.34
	AT 融合	83.12±1.01	84.99±0.22
LF	ADD-Fusion	83.61±0.46	84.80±0.41
	拼接融合	83.54±1.77	84.43±0.55
	AT 融合	83.20±1.72	84.77±0.63
LGF	ADD-Fusion	83.39±0.54	**85.40**±0.40
	拼接融合	83.23±1.57	84.60±0.47
	AT 融合	81.79±1.37	84.52±0.72

表2。IEMOCAP数据集上采用不同融合方法和融合层的多模态模型的 WA(%) 结果。FL 表示融合层，FM 表示融合方法。WA(%)$\mathcal{L}_s$ 表示使用单一损失 $\mathcal{L}_s$ 训练模型，而 WA(%)$\mathcal{L}_p$ 表示使用视角损失 $\mathcal{L}_p$训练模型。“±”之后的数字表示标准差。

权重重复实验20次，并报告平均结果以降低随机性。我们选择加权准确率（WA） [2]作为评估指标，用于将我们的方法与以往方法进行比较。

4. 结果与分析

本节展示多模态性、不同融合方式和视角损失对语音情感识别的影响。随后，我们将所提出模型的结果与其他最先进的方法进行比较。

4.1. 单模态与多模态

表1和表2分别展示了我们的单模态框架和多模态框架在 IEMO-CAP 数据集上的结果。如表1所示，文本模态的表现显著优于音频模态。这是合理的，因为我们发现，大多数语音情绪都可以直接从文本中推断出来。融合音频和文本模态信息后，多模态模型在大多数设置下都优于单模态模型。这一现象表明，两种模态中的信息可以互为补充，也为设计各种多模态融合模型以提升语音情感识别效果提供了可能。

4.2. 不同融合方式的效果

如表2所示，在比较不同融合方法时，我们发现，无论在哪些位置执行融合操作，ADD-Fusion 在这三种方法中表现最佳。这可能是因为 ADD-Fusion 操作平等地利用两种模态，使它们在特征

		IEMOCAP上的情绪（%）			
模型	模态	开心	sad	中性	愤怒
单模态	A	72.01	75.51	60.16	88.24
	T	90.52	90.61	71.88	75.88
单一损失	A	48.76	86.53	60.51	84.12
	T	84.20	71.43	78.12	65.88
	A+T	89.62	88.57	73.44	81.18
视角损失	A	58.24	73.88	59.64	90.59
	T	90.29	88.98	76.82	61.18
	A+T	90.52	89.39	76.56	88.24

表3。ADD-Fusion 在 IEMOCAP 不同情绪上的 LGF 处，单模态模型和多模态模型的准确率。A 和 T 分别指音频和文本模态。

方法	**WA(%)**
bc-LSTM（2017年）[10]	75.60†
CATF-LSTM（2017年）[1]	80.10†
Zheng.（2019年）[20]	78.02
DANN（2020年）[21]	82.68
CONSK-GCN（2021年）[14]	84.79‡
Soumya.（2022年）[18]	83.80
我们的方法	**85.40***

表4。IEMO-CAP 数据集上不同方法的性能。 † 表示音频+文本模态的结果。
‡ 表示对开源代码进行10次复现所得的平均结果。我们的方法是采用 LGF+ADD-Fusion+视角损失的多模态模型。 *表示我们的方法显著优于其他方法（p-值 < 0.01）。

空间中更好地互补。相比之下，拼接融合和 AT 融合可能会因加权组合而忽略某些模态。

在不同层进行融合时，使用单一损失的早期融合（EF）搭配 ADD-Fusion 表现更好，这表明早期融合能让多模态信息更充分地交互。相比之下，使用视角损失来增强模态特征时，LGF 的表现最佳。我们推测，视角损失使 LGF 中的特征彼此解耦，从不同视角促进情感识别。

4.3. 单一损失与视角损失

从表2可以看出，在大多数融合设置下，采用视角损失的多模态模型在性能上有明显提升。为理解这一点，我们以采用 ADD-Fusion 的 LGF 为例，探究视角损失效果显著的原因。表3展示了不同模态和模型在不同情绪上的详细结果。

对于单模态结构，文本模态对开心和悲伤情绪的识别能力出色，而音频模态在识别愤怒情绪方面表现良好。这是合理的，因为用户通常会使用明确的措辞表达开心和悲伤情绪，因此文本模态对这两类情绪的预测效果更好。相比之下，根据用户的语气判断其是否愤怒较为容易，而这种情绪通常体现在音频信号中。

对于采用单一损失的多模态模型，上述三种情绪的识别性能均有所下降。一个可能的原因是，采用简单融合时，文本模态主导了信息，而音频模态的作用被削弱。具体而言，多模态模型（A+T）在愤怒情绪上的性能比单模态音频框架大幅下降，最高达 7%。此外，该模型在开心和悲伤情绪上的表现也略有下降，因为简单融合会混合多模态特征，并掩盖一些关键的单模态信息。

使用视角损失后，多模态模型在进行融合时能够保留这些特定情绪的关键判别性单模态特征。我们观察到模型在开心、悲伤和愤怒情绪上的表现相同或相近。此外，模型在中性情绪上的表现也有所提升。最后，两种模态发挥着不同的作用，共同提升了整体表现。

4.4. 所提方法与以往工作的比较

表4展示了不同方法在 IEMOCAP 数据集上的性能。相比之下，我们的方法（LGF+ADD-Fusion+视角损失）比此前最先进的方法提升了 0.61%。这一结果表明，我们的方法简单却非常有效，适用于语音情感识别。

实际上，以往的工作也使用音频和文本模态特征。然而，它们通常在没有视角损失约束的情况下过早融合多模态特征，这可能会引入一些与情绪无关的噪声信息。相比之下，LGF 利用了具有判别性的情感信息。此外，视角损失使每种模态在情绪预测中都能发挥自身特点。从表2和表4可以看出，不使用视角损失的多模态模型不如最先进的方法。视角损失显著促进了情绪预测。

5. 结论

在这项工作中，我们通过实证研究了基于音频和文本的语音情绪识别，并探讨了不同模态融合方式的影响。我们发现，ADD-Fusion 是一种简单而有效的方法。在此基础上，我们进一步提出了视角损失，以在融合时保留特定情绪的关键且具有区分性的单模态特征。结果表明，我们的方法在 IEMOCAP数据集上达到了最先进的性能。我们认为，这项研究有助于其他关于多模态模型和任务的研究。

6. 参考文献

[1] Soujanya Poria、Erik Cambria、Devamanyu Hazarika、Navonil Mazumder、Amir Zadeh 和 Louis-Philippe Morency，“用于情境多模态情感分析的多层多重注意力”，载于*ICDM*，2017年，第 1033–1038 页。

[2]Seyedmahdad Mirsamadi、Emad Barsoum 和 Cha Zhang，“结合局部注意力的循环神经网络用于自动语音情感识别”，载于*ICASSP*，2017年，页码 2227–2231。

[3] Yuya Chiba、Takashi Nose 和 Akinori Ito，“用于语音情感识别的基于多流注意力的 BLSTM 与特征分段”，载于*Interspeech*，2020年，页码 3。

[4] Yuan Gao、JiaXing Liu、Longbiao Wang 和 Jian wu Dang，“基于注意力的特征级融合域对抗自编码器用于语音情感识别”，载于*ICASSP*，2021年，页码 6314–6318。

[5] DN Krishna 和 Ankita Patil，“使用跨模态注意力和一维卷积神经网络进行多模态情感识别”，载于 *Interspeech*，2020年，页码 4243–4247。

[6] Haiyang Xu、Hui Zhang、Kun Han、Yun Wang、Yip- ing Peng 和 Xiangang Li，“从语音中学习多模态情绪识别的对齐”，载于 *Interspeech*，2019年，第3569–3573页。

[7] Yoonhyung Lee、Seunghyun Yoon 和 Kyomin Jung，“使用跨注意力和对齐的音频与文本进行多模态语音情感识别”，载于 *Interspeech*，2020年。

[8]Seunghyun Yoon、Seokhyun Byun、Subhadeep Dey 和 Kyomin Jung，“使用多跳注意力机制进行语音情感识别”，载于 *ICASSP*，2019年，第2822–2826页。

[9] Seunghyun Yoon、Subhadeep Dey、Hwanhee Lee 和 Kyomin Jung，“用于语音情感识别的注意力模态跳转机制”，载于 *ICASSP*，2020年，第3362–3366页。

[10] 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》，发表于*ACL*，2017年，第873–883页。

[11] Sepp Hochreiter 和 J¨urgen Schmidhuber，“长短期记忆”，载于*Neural computation*，1997年，第 1735–1780 页。

[12] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin，“注意力就是你所需要的一切”，载于*NIPS*，2017年，第 5998–6008 页。[13] Dou Hu、Xiaolong Hou、Lingwei Wei、Lianxin Jiang 和 Yang Mo，“MM-DFN：用于对话情感识别的多模态动态融合网络”，载于*ICASSP*，2022年，第 7037–7041 页。

[14]Yahui Fu、Shogo Okada、Longbiao Wang、Lili Guo、Yaodong Song、JiaXing Liu 和 Jianwu Dang，《CONSK-GCN：用于多模态情感识别的对话语义与知识导向图卷积网络》，发表于*ICME*，2021年，第1–6页。

[15]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》，发表于语言资源与评估，2008年，第42卷，第335–359页。

[16] Wenxiang Jiao、Haiqin Yang、Irwin King 和 Michael R Lyu，《用于话语级情感识别的分层门控循环单元（HiGRU）》，载于 *NAACL-HLT*，2019年，第397–406页。

[17] Zhen Wu、Xin-Yu Dai、Cunyan Yin、Shujian Huang 和Jiajun Chen，《通过用户注意力和产品注意力改进情感分类的评论表示》，载于*AAAI*，2018年，第32卷。

[18] Soumya Dutta 和Sriram Ganapathy，《采用可学习前端和自注意力进行情感识别的多模态变换器》，载于 *ICASSP*，2022年，第6917–6921页。

[19]Devamanyu Hazarika、Soujanya Poria、Amir Zadeh、Erik Cambria、Louis-Philippe Morency 和 Roger Zim- mermann，《用于双人对话视频情感识别的对话记忆网络》，载于 *NAACL*，2018年，第2122页。

[20] Zheng Lian、Jianhua Tao、Bin Liu 和 Jian Huang，“通过注意力机制进行对话情感分析”，发表于*Interspeech*，2019年。

[21] Zheng Lian、Jianhua Tao、Bin Liu、Jian Huang、Zhanlei Yang 和 Rongjun Li，“用于多模态情绪识别的上下文相关领域对抗神经网络”，发表于*Interspeech*，2020年，第394–398页。