HCAM——用于多模态情绪识别的分层交叉注意力模型

Soumya Dutta, *IEEE* 学生会员，以及 Sriram Ganapathy，*IEEE* 高级会员

arXiv:2304.06910v2 [eess.AS] 9 Jan 2024

摘要——**由于情绪表达具有多模态特性，对话中的情绪识别颇具挑战。我们提出一种结合循环神经网络模型与协同注意力神经网络模型的分层交叉注意力模型（HCAM）方法，用于多模态情绪识别。模型输入包含两种模态：i）音频数据，经可学习的 wav2vec 方法处理；ii）文本数据，使用双向 Transformer 编码器表示（BERT）模型进行表征。音频和文本表示经由一组带有自注意力的双向循环神经网络层处理，将给定对话中的每个话语转换为固定维度的嵌入表示。为纳入上下文知识以及两种模态间的信息，我们使用协同注意力层组合音频和文本嵌入，使其能够对与情感识别任务相关的话语级嵌入赋予不同权重。音频层、文本层以及多模态协同注意力层中的神经网络参数均针对情感分类任务进行分层训练。我们在三个成熟的数据集 IEMOCAP、MELD 和 CMU-MOSI 上进行了实验，结果表明，所提出的模型显著优于其他基准模型，并有助于在所有这些数据集上取得最先进的结果。**

索引词—**分层学习、协同注意力模型、多模态融合、情绪识别。**

一、引言

HUMAN 以复杂的多模态方式表达的情绪，在人际互动和人际交流中发挥着核心作用。随着自动化系统在日常生活中无处不在，对人类情绪的理解已成为设计这些系统的关键环节。能够执行情感识别的机器有助于开发个性化的人机交互界面，例如对话代理 [1],社交媒体分析 [2],客户呼叫中心 [3] 以及心理健康监测系统 [4]。这项关键任务被称为对话中的情绪识别（ERC）。

对话数据中的情绪识别面临诸多挑战，具体包括多个（重叠的）说话人、短期和长期依赖关系、 [5],轮次事件持续时间较短、音频中的背景噪声等。此外，情绪识别任务本质上是多模态的，信息可以通过多种方式表达，例如面部表情、 [6],语音、 [7], 手势、 [8], 生理信号、 [9] 或通过一种……

S. Dutta 和 S. Ganapathy 隶属于印度班加罗尔 560012 印度科学研究所电气工程系的学习与提取及声学模式（LEAP）实验室。本研究获得了印度教育部总理研究奖学金（PMRF）项目提供的资助。电子邮箱：{soumyadutta, sriramg} iisc.ac.in。稿件收到日期：2023 年 3 月 xx 日。

这些模态的组合。不同模态包含与潜在情绪相关的不同程度的信息，因此，设计一种用于情绪识别的联合多模态方法通常被认为是必要的 [10]。

以往的研究探索了将音频与视觉信号 [11], [12] 以及文本与音频信号 [13],[14]结合使用。此外，也有研究尝试使用全部三种模态，即视觉、音频与文本 [15]。虽然需要具备以多模态方式感知情绪的能力，但对于某些模态的数据不可用或存在噪声的情况，也有必要稳健地识别各个模态中的情绪。

本文针对多模态情绪识别问题，提出了一种新颖的分层交叉注意力模型（HCAM）。我们的建模框架涉及基于语音和文本的大规模表征学习网络，并采用层次训练过程。该训练范式由三个不同的阶段组成，其中第一阶段以无上下文方式训练单模态预测器。作为对话情感识别中不可或缺的组成部分，上下文信息随后在第二阶段被引入。最后一个阶段则通过协同注意力模块融合声学与文本流。音频特征采用 wav2vec 模型 [16] 提取，而文本表示则通过 RoBERTa [17] 模型获得。协同注意力是一种联合建模听觉信息与文本信息的机制。借助基于音频和基于文本的嵌入表示之间天然的对称性，这种融合技术引导模型关注用于情感识别的多模态上下文信息。为了以层次化方式训练该模型，我们探索了监督式对比损失 [18]的应用。

实验在三个成熟的数据集上进行，分别为 IEMOCAP [19],、MELD [20] 和 CMU-MOSI [21]。尽管不同数据集中的情绪类型、话语时长、对话长度、音频中的噪声或失真程度以及对话风格各不相同，但本文提出的 HCAM 架构适用于所有这些数据集。

本研究的主要贡献如下：

- 我们提出了一种用于多模态情绪识别的分层方法：首先分别在各模态中对话语级信息进行处理，然后进行话语间对话建模，最后通过交叉注意力进行多模态处理。实验结果表明，分层建模对于提升情感识别性能至关重要。

- 由于模型在每个阶段学习对话的不同方面，我们提出在测试集期间结合之前训练阶段和当前阶段的模型预测，以进一步提升性能

- 在模型训练过程中，除了交叉熵损失外，我们还使用了监督式对比损失。监督式对比损失使模型能够更加关注难以学习的训练样本。

- 我们在三个基准数据集上测试了所提出的模型，并在所有这些数据集上取得了最先进的结果。此外，我们还尝试在测试阶段使用自动语音识别转录文本替换真实文本，以测试模型的鲁棒性。

II. 相关工作

A. 音频特征提取

情感识别的早期方法主要采用基于知识的特征，例如振幅、基频和频谱轮廓，这些特征由 Sauter 等人提出。 [22] Luengo 等人将韵律参数用于语音情感识别，而 Koolagudi 等人则使用了片段级韵律特征。 [23] [24] Eyben 等人指出，梅尔频率倒谱系数能够带来最佳的情感分类表现。 [25] 最近，Schuller 等人在 Interspeech 语言旁信息挑战赛中发现，若干基于知识的特征的统计描述符在语音情感识别任务中表现显著更优。 [26] 这些特征经过进一步优化， [27] 形成了一组精简的参数（openSMILE工具包）。 [28]

近年来深度学习的发展促使研究人员开发本质上具有可学习性的音频特征提取器，例如Ravanelli等人提出的SincNet特征、Zeghidour等人提出的 [29], 可学习音频前端（LEAF），以及Agrawal等人提出的[30] 可解释高斯滤波器 [31]，[32]。最近，研究人员利用大规模未标注语音数据语料库提出了无监督和自监督方法 [33], [16]。近期的多项情感识别研究探索了这些表示的应用 [34], [35], [36]。

B. 文本特征提取

早期的文本特征提取方法采用词袋法，并结合词语关系特征，正如 Xia 等人提出的那样 [37]。受深度学习启发，近期的模型利用预测任务来学习文本嵌入。Mikolov 等人提出的一种此类方法被称为 word2vec 模型 [38], , 广泛应用于文本情感分析的特征提取 [39], [40], [41]。随着循环网络和注意力网络的发展，例如双向长短期记忆网络（双向长短期记忆网络）、门控循环单元（门控循环单元）和 Transformer [42],, 研究人员开发出了 BERT [43] 和 RoBERTa [17] 等改进的语言模型。BERT嵌入在情感分析 [44]等多种下游任务中取得了性能提升。

C. 多模态融合

多种模态的融合已被证明有助于情感识别 [45]。早期尝试采用表示拼接的方法已被多模态 Transformer 等先进技术取代 [46]，[47]。Zadeh 等人提出了一种多视图序列学习架构 [48]。随后，研究者利用动态融合图对其进行了改进 [49]。基于注意力的融合方法也被应用于各种视觉-语言任务 [50]，[51]。

D. 在情感识别对话中融入上下文信息

对话中存在多个说话人可能会产生短期和长期依赖关系。正如 Poria 等人 [5] 所指出的，这给情感识别对话带来了重大挑战。为了解决这一问题，Poria 等人 [45] 使用长短期记忆网络来捕获对话中的话语间上下文。随后，Majumder 等人 [52] 在其名为 DialogueRNN 的模型中进一步改进了这一方法。Ghosal 等人的研究提出了使用图卷积网络 [53] 的方法。

E. 与先前研究的对比

与以往工作相比，所提出的 HCAM 框架在以下方面具有创新性。

- 我们提出了一种基于课程学习的建模阶段设计方案。在这一设计中，首先学习较为简单的任务，即识别对话中话语级别的情感状态。随后，在基于循环层的话语级模型之上，设计更为复杂的话语间上下文建模任务。此外，基于神经注意力的建模层实现了多模态整合。该分层建模框架能够从语音和文本表示中高效地学习潜在的情绪标签。

- 本文探讨了用于情感识别任务的监督式对比损失。该损失通过关注较难的训练样本，提升了模型的学习能力。

- 模型从对话中识别情绪的能力会随建模阶段而变化。我们提出在推理期间集成不同阶段的模型预测。

III. 背景

A. wav2vec

wav2vec 是一种基于自监督原理的表示学习框架 [33]。在 wav2vec 2.0 模型 [16]，中，音频信号被切分为短小且相互重叠的帧。每个切分后的片段都会经过卷积特征提取层，随后通过量化模块将表示编码到离散空间中。这些表示通过变换器编码器层进行上下文化。该网络采用自监督对比损失进行预训练。

B. RoBERTa

近年来，Devlin为构建大规模语言模型做出了重大贡献之一。

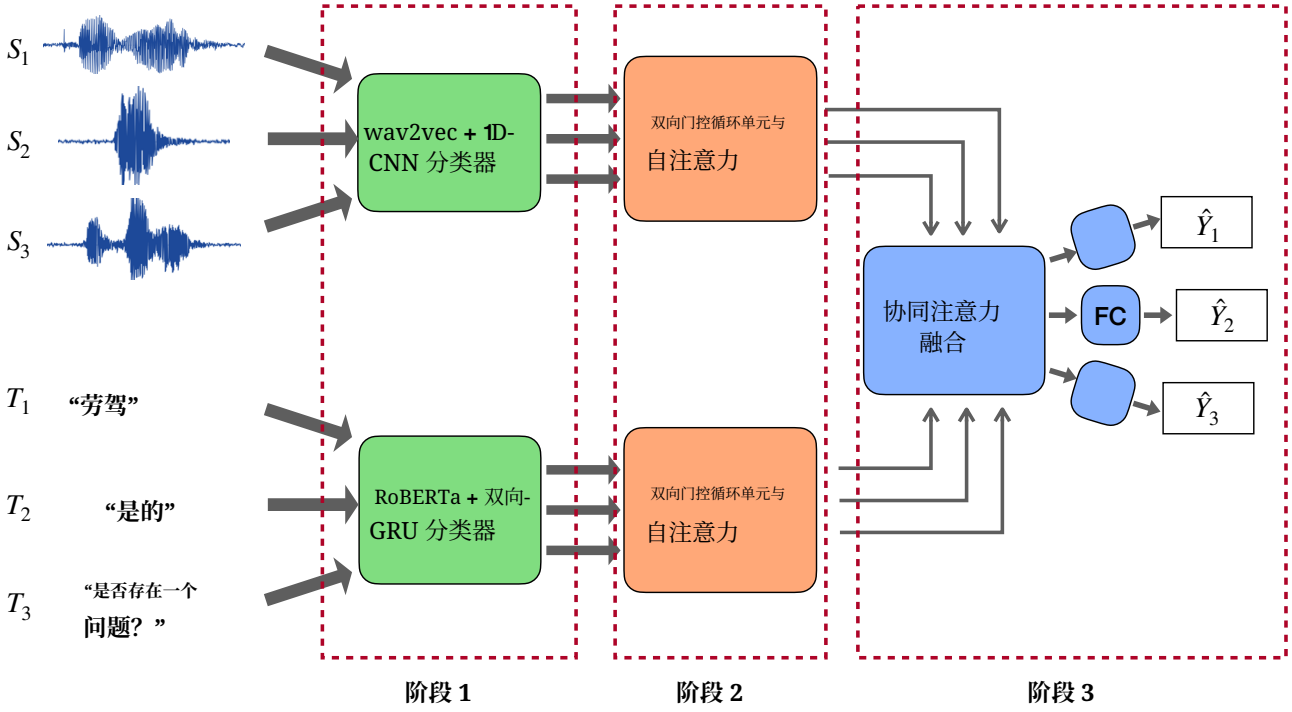


图1：所提出模型的框图。这里，S_1、S_2 和 S_3 指对话中的语音话语。同样，$\hat{}\hat{}\hat{}$ T_1、T_2 和 T_3 指相应语音信号的文本转录。Y_1、Y_2 和 Y_3 指三个话语的预测情绪标签。图中还标出了训练的三个阶段。

等人 [43] 在构建大规模语言模型方面作出了重要贡献。这种架构称为来自 Transformer 的双向编码器表示（BERT），在文本语料库上通过两项任务进行训练：预测句子中被遮蔽的词（称为掩码语言建模），以及预测两个句子在语义上是否前后相接（称为下一句预测）。Liu 等人 [17] 在移除下一句预测任务后，使用规模更大的文本语料库训练了这一架构。这种预训练模型称为稳健优化的 BERT 方法（RoBERTa），还对模型设计中的若干其他超参数进行了优化。

C. 自注意力与交叉注意力

Vaswani 等人提出的自注意力机制 [42]，处理一个长度为 N、维度为 d_k 的序列。该序列被转换为三个矩阵：$Q \in \mathcal{R}^{N\times d_k}$、$K \in \mathcal{R}^{N\times d_k}$ 和 $V \in \mathcal{R}^{N\times d_k}$。自注意力通过计算查询表示（记为 Q）与键表示（记为 K）之间的相似度来实现。该相似度矩阵经 Softmax 函数转换为概率分布后，用于对值表示（记为 V）进行加权求和。最后，将查询矩阵（Q）与该加权和相加，再经过一个层归一化块。

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$
$$Q = Q + Attention(Q,K,V) \quad (2)$$
$$output = LayerNorm(Q) \quad (3)$$

交叉注意力网络与自注意力模块类似，但有一个关键区别：查询矩阵和键／值矩阵由不同模态的表示构成。如果我们将音频模态中的查询矩阵（记为 Q_A）与文本模态中的键／值矩阵（记为 K_T/V_T）作为输入，则交叉注意力网络的操作如下：

$$Atten(Q_A,K_T,V_T) = softmax(\frac{Q_A K_T^T}{\sqrt{d_k}})V_T \quad (4)$$
$$Q_A = Q_A + Atten(Q_A,K_T,V_T) \quad (5)$$
$$Q_A = LayerNorm(Q_A). \quad (6)$$

IV. 所提出的架构

所提出模型的框图如图1所示。由于整个模型的架构相对庞大，由文本和音频两种模态驱动，因此我们的模型包含三个不同的阶段，这些阶段采用分层方式进行训练，其中每个阶段均使用前一阶段的预训练模型参数，且不进行微调。

所有阶段均采用层次化方式进行训练，使用交叉熵损失与监督式对比损失函数（第 IV-D 节）的加权组合。模型实现已公开提供[1]。

[1]https://github.com/iiscleap/多模态情感识别，适用于协同注意力

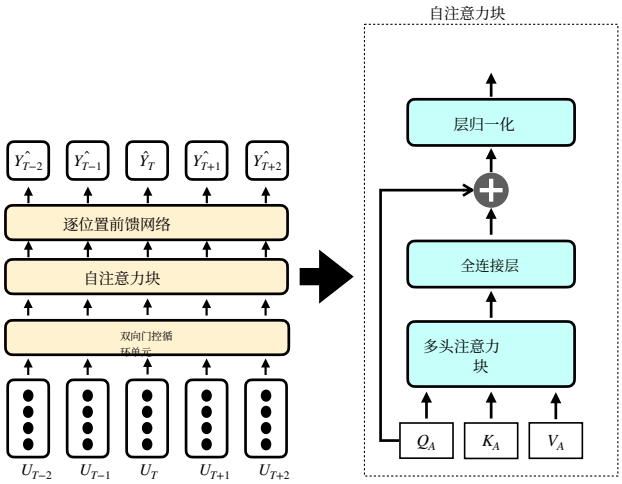


图2：带有自注意力机制的上下文GRU框图。这里， U_T、$U_{T\pm1}$和 $U_{T\pm2}$指模型阶段I中对话内每个话语的单模态嵌入。

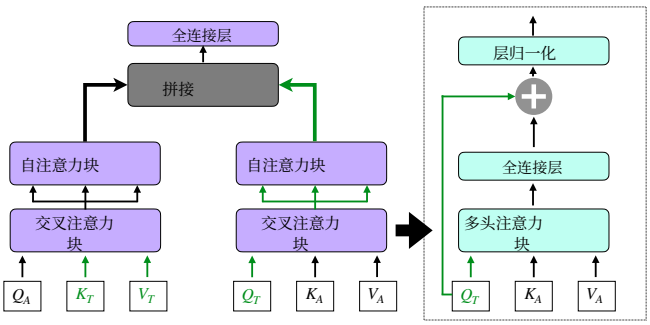


图3：所提出模型中使用的协同注意力网络。它由两个子块组成——交叉注意力块和自注意力块。

*A. HCAM*阶段*I*

在第一阶段，我们使用音频和文本训练话语级嵌入表示提取器。训练模型对单条话语进行分类，而不考虑话语之间的对话上下文。在此阶段训练的模型根据音频信号或文本转录中包含的线索，将单条话语分类到相应的情绪类别中。预训练特征提取器通常是计算需求较高的模型，这限制了每次迭代中可处理的话语数量。单次对话中的大量话语使以往研究难以微调这些特征提取器，因为它们将对话作为整体进行处理。我们的分层建模使我们能够在此训练阶段微调预训练特征提取器生成的嵌入，从而提升对每条单独话语的情感分类效果。

1）音频嵌入提取器：将音频输入在 Libri-light [55], CommonVoice [56], Switchboard [57] 和 Fisher [58] 数据集上预训练的 wav2vec2.0 large模型 [16], 。随后，使用 Switchboard 语料库中 300 小时的嘈杂电话对话数据对模型进行微调。受 Pepino 等人提出的策略启发 [36],，我们对 wav2vec2.0 网络中的 Transformer层进行微调，同时保持较低层的卷积层不变。将 wav2vec 模型所有 Transformer层的隐藏层输出在帧级求和，然后输入一维 CNN网络。最后，对 CNN 网络输出的嵌入在话语级进行平均池化，为给定话语生成音频嵌入。将 1D-CNN 网络输出的嵌入输入上下文GRU层。

2）文本嵌入提取器：文本数据的嵌入提取器采用了与音频特征提取类似的架构。我们通过预训练的 RoBERTa 模型 [17] 获取词嵌入，并将其拼接起来。

模型最后四层隐藏层的表示。双向门控循环单元（bi-GRU）层能够整合话语内部的上下文信息，从而生成话语级嵌入表示。与音频嵌入表示的提取不同，RoBERTa 和 bi-GRU 模型的所有层均经过微调。

*B. HCAM*阶段*II*

上一建模阶段获得的话语级嵌入表示在本阶段中被使用，在此阶段，我们通过带有自注意力机制的双向门控循环单元（双向门控循环单元）架构引入话语间上下文。阶段I中在每个单独话语上训练的模型所提取的表示，进一步结合了对话上下文信息。

1) 话语间上下文*GRU*：我们提出了一个简单的块，称为上下文GRU，它会考虑对话中所有话语的信息。上下文GRU的框图如图2所示。尽管双向门控循环单元本身能够整合上下文信息，但对于话语数量较多的对话，它可能无法捕获长期依赖关系。因此，我们采用自注意力来有效建模长期上下文。自注意力块的输出经由带有ReLU激活的逐位置前馈层进行处理。

C. HCAM 第三阶段

模型的第三阶段由不同模态的嵌入表示的有效融合构成。

1) 多模态融合：协同注意力块的网络架构如图3所示，由两个子块组成，即交叉注意力和自注意力。自注意力和交叉注意力机制已在第III-C节中讨论。在协同注意力网络中，我们利用不同模态之间的交叉注意力。音频与文本之间的交叉注意力有两种实现方式：一种是从音频数据中获取查询表示，而从文本数据中获取键和值表示；另一种则是将音频与文本的角色互换。交叉注意力块输出的跨模态嵌入会通过添加一个进一步得到增强。

自注意力块。交叉注意力的两个分支如图3所示，拼接后形成多模态表示。

融合音频与文本信息的最终表示将通过位置前馈层。

D. 损失函数

1）跨对话的监督式对比损失：我们探讨由 Khosla 等人提出的监督式对比损失函数 [18]，，该函数促使来自同一情绪类别的话语表示彼此相似。设小批量大小为 $\mathcal{B}$，其中从多个对话中得到的话语级表示记为 $\{x_1, x_2, \ldots, x_{\mathcal{B}}\}$。损失公式中的特征均经过归一化，即 $||x_i|| = 1 \quad \forall i = \{1, 2, \ldots, \mathcal{B}\}$。我们将相应的标签记为 $\{y_1, y_2, \ldots, y_{\mathcal{B}}\}$。对于索引为 j 的样本，小批量中的正样本集合记为P_j： $\{i \in \mathcal{B} \quad s.t. \quad y_i = y_j\}$。监督式对比损失为:

$$L^{sup-con} = \sum_{j \in \mathcal{B}} \frac{-1}{|P_j|} \sum_{p \in P_j} log \frac{exp(x_j^T x_p / \tau)}{\sum\limits_{a \in \mathcal{B}} exp(x_j^T x_a / \tau)}, \quad (7)$$

其中， τ 是一个超参数，表示该损失的温度。通常，对比损失用于表示学习任务。对比损失依赖真实标签，因此也可以将其用于监督场景。除了交叉熵损失外，该损失还会使系统重点关注难分类样本。然而，与表示学习任务不同，ERC 数据集规模不够大，无法仅凭对比损失对话语进行分类。因此，我们在训练情绪识别模块时，将监督式对比损失与交叉熵损失结合使用。

2) 组合损失：我们采用交叉熵损失与监督式对比损失的凸组合。用于训练模型所有阶段的最终损失定义如下:

$$L^{tot} = \beta CE(y, \hat{y}) + (1 - \beta) L^{sup-con}, \quad (8)$$

其中， β 是一个取值范围为 $[0, 1]$的超参数。

E. 推理

在特定阶段的推理过程中，我们会结合前一阶段模型的预测结果。因此，一旦上下文GRU模型在阶段II中完成训练，我们就会结合阶段I中音频模型和文本模型的预测结果。在第三阶段的推理过程中，音频和文本表示经过融合后，我们会结合上下文GRU模型的预测结果（分别针对音频和文本）。设话语 x 在阶段I结束时的 Softmax 输出分别为 p_a^1 （音频）和 p_t^1 （文本）。这些输出将与单模态上下文GRU预测（分别用 p_a^2 和 p_t^2表示），其权重分别用 $\alpha_a^{1\to2}$和 $\alpha_t^{1\to2}$表示，如下所示，

$$\hat{y}_a^2(x) = \alpha_a^{1\to2} p_a^2(x) + (1 - \alpha_a^{1\to2}) p_a^1(x) \quad (9)$$

$$\hat{y}_t^2(x) = \alpha_t^{1\to2} p_t^2(x) + (1 - \alpha_t^{1\to2}) p_t^1(x) \quad (10)$$

同样，我们将第二阶段的输出与协同注意力模块的预测结果（p_c）结合起来，如下所示:

$$\hat{y}(x) = \alpha_c p_c(x) + \alpha_a^{2\to3} \hat{y}_a^2(x) + \alpha_t^{2\to3} \hat{y}_t^2(x) \quad (11)$$

其中，$\hat{}y(x)$ 指用于分类的最终预测结果。公式（9）、（10）和（11）中使用的所有组合权重均根据验证集上的性能确定。

V. 实验与结果

A. 数据集

我们在三个广泛使用的数据集上评估了所提出的方法，分别为 IEMO-CAP [19], MELD [20] 和 CMU-MOSI [21]。与前两个数据集不同，CMU-MOSI 本质上并非多说话人数据集，因为它包含单人独白。此外，本文仅使用文本和音频模态完成情感识别任务。

1) IEMOCAP： IEMOCAP 数据集由 151 段视频录制组成，分为 5 个场次。每个场次都是由一男一女两人参与的一段对话。每段录音又被划分为多个话语。总共有 10,039 个话语，每个话语均由人工标注员标记为 10 种情绪之一：愤怒、快乐、悲伤、中性、沮丧、兴奋、恐惧、惊讶、厌恶或“其他”。与以往的研究保持一致，我们采用四分类任务，考虑愤怒、快乐、悲伤、中性和兴奋这几种类别（其中兴奋和快乐类别合并）。四种情绪标签共包含 5531 个话语。此外，我们还采用了另一种包含 6 个情绪类别的设置，这与部分先前研究（如 [52]）的做法一致。选取前 6 种情绪类别，共得到 7433 个话语。该数据集存在类别不平衡问题，其中快乐情绪的样本数量最少（训练集中的话语数量为 648）。对于这两种设置，我们均使用场次 5 进行测试。我们选择场次 1 用于验证模型，并使用场次$2-4$进行训练。

2) MELD：MELD 数据集是一个多方数据集，基于热门电视剧《老友记》的视频片段构建。训练集包含 9988 条话语，验证数据包含 1108 条话语，测试数据包含 2610 条话语。该数据集执行七分类任务，每条话语均标注为 7 种情绪之一：愤怒、悲伤、喜悦、中性、恐惧、惊讶或厌恶。与 IEMOCAP 六分类问题类似，该数据集也存在类别不平衡现象，其中中性是占比最高的类别标签，厌恶则是出现频率最低的类别标签（训练集中的话语数量分别为 4710 条和 271 条）。

3) CMU-MOSI：CMU-MOSI 数据集共包含 93 段独白，划分为 2199 条话语。每段独白又被划分为若干条话语，并按照 $[-3, 3]$ 的范围进行标注。遵循以往的研究工作，我们将其视为一个二分类问题，其中情感值处于 $[-3, 0)$ 范围内的话语被归类为负面情感，而

表 I：不同数据集上的加权 F1 分数结果。

模态（训练集阶段）	IEMOCAP		MELD	CMU-MOSI
	4 路	6 种方式		
音频（阶段I）	64.3%	55.6%	48.2%	64.4%
音频（第二阶段）	78.7%	65.7%	50.1%	67.4%
文本（阶段I）	68.4%	53.8%	63.3%	84.3%
文本（第二阶段）	81.4%	64.4%	65.6%	85。4%
音频 + 文本（第三阶段）	85.9%	70.5%	65.8%	85.8%

将取值范围为 $[0,3]$ 的样本视为正面情感。在数据集划分方面，我们遵循 Poria 等人 [45], 的先前研究，使用前 62 段独白进行训练和验证，最后 31 段独白用于测试。在这 62 段独白中，我们使用其中 49 段训练模型，其余 13 段用于验证。

B. 实现细节

模型使用 Adam 优化器进行训练，学习率为 $1e-5$，在 MELD 和 IEMOCAP 数据集上的批量大小为 32，而在 CMU-MOSI 数据集上的批量大小则减小至 8。本研究报告的所有实验均采用 5 种随机权重初始化方案。以下所有实验均报告随机初始化下的平均性能。我们将模型的不同阶段训练 100 轮，因为我们发现验证性能在此范围内趋于饱和。在所有实现中，我们还使用了 L2 范数为 0.25 的梯度裁剪。

C. 结果

我们报告了所提出模型在各个单独模态下的性能，以及该模型在多模态设置下的性能。三个数据集上的关键结果如表 I 所示。

The followi[ng are] the ob[serva] tions from th[这些结果]lt s，（i）在 IEMOCAP 数据集中，音频和文本模态的表现相对接近，而在 MELD 和 CMU-MOSI 数据集中，音频的表现逊于文本领域。（ii）上下文添加框架在所有三个数据集中均证明了其有效性。所提出的结合自注意力的上下文 GRU 架构虽然结构简单，但能够提升两种模态的性能。对于音频模态，与阶段 I 的性能相比，IEMOCAP 四分类、IEMOCAP 六分类、MELD 和 CMU-MOSI 数据集的相对提升幅度分别为 40.3%、22.7%、3.7% 和 8.4%。类似地，对于文本模态，我们观察到其相对提升幅度分别为 41.1%、22.9%、6.3% 和 7%。在这三个数据集中，IEMOCAP 的每段对话包含的话语数量最多，而 MELD 的话语数量最少。随着数据集中话语数量的增加，上下文建模的效果也越来越显著。

(iii) 最终的多模态融合在三个数据集上的表现均优于任何单一模态。

表 II：与其他 IEMOCAP 四分类研究的比较。所有分数均为加权 F1 分数。

系统	音频	Text	音频+文本
Majumder 等人 [39]	57.1%	73.6%	76.1%
Mai 等人 [59]	38.2%	**81.5**%	80.6%
Li 等人 [60]	69.3%	-	79。1%
Mittal 等人 [15]	-	-	82.4%
Mai 等人 [61]	-	-	82。5%
Dutta 等人 [47]	73。5%	78。9%	83。7%
Lian 等人 [44]	-	-	84。8%
HCAM（本研究）	**78**。**7**%	81.4%	**85**。**9**%

表 III：IEMOCAP 六分类任务与其他研究成果的比较。所有分数均为加权 F1 分数。* 表示我们的实现。

系统	音频	Text	音频+文本
Majumder 等人 [52]	-	-	62.8%
Ghosal 等人 [53]	-	-	64.2%
Shen 等人 [62]	-	-	65。9%
Mao 等人 [63]	-	-	69.7%
Lian 等人 [44]	-	-	70.5%
Li 等人 [64]	-	-	69.3%*
HCAM（本研究）	65.7%	64.4%	**70.5**%

模态。与 IEMOCAP 四分类、IEMOCAP 六分类、MELD 和 CMU-MOSI 数据集上表现最佳的单一模态相比，多模态结果分别取得了 24.2%、14%、0.6% 和 2.7% 的相对提升。这些结果表明，即使某些模态在情感分类任务中的表现不如其他模态，协同注意力机制仍能够有效提升模型性能，使其优于表现最佳的单一模态。

VI. 讨论

A. 与其他工作的比较

我们对本研究与文献中的其他现有研究进行了比较。表 II 展示了在 IEMOCAP 数据集上进行四分类时与相关研究的比较。可以看到，我们的模型在音频信号上的性能显著优于其他研究（相对提升约 20%）。这归功于对 wav2vec 模型生成的音频表示进行高效的上下文建模。

虽然我们未能在文本模态上取得最先进的结果，但在多模态融合任务中，我们以 7% 的相对优势超越了此前的最佳结果。我们还将本模型与此前在 IEMOCAP 数据集上报告六分类结果的其他研究进行了比较（表 III）。对于这种类别不平衡的分类设置，我们发现本模型的结果与 Lian 等人 [44] 报告的结果相当。

我们在表 IV 中将本文与 MELD 数据集上的其他相关研究进行了比较。与 IEMOCAP 类似，我们发现仅使用音频时的性能有了显著提升。音频信号性能较当前最先进的结果相对提升了约 14%。仅使用文本模态时，我们模型的性能也有所提升。

表IV：与 MELD 数据集上已有研究的比较（以加权 F1 分数衡量）。

系统	音频	Text	音频+文本
Poria 等人 [45]	39。1%	56.4%	59.3%
Majumder 等人 [52]	41.8%	57%	60.3%
Zhang 等人 [65]	42。2%	57.4%	59.4%
Ghosal 等人 [53]	-	58.1%	-
Li 等人 [66]	-	61.9%	-
Shen 等人 [62]	-	62.4%	-
Mao 等人 [63]	-	-	63.6%
Lian 等人 [44]	-	-	64.5%
Li 等人 [64]	-	-	64%
HCAM（本研究）	**50.1**%	**65.6**%	**65.8**%

表 V：CMU-MOSI 数据集与其他研究成果的比较。所有分数均为加权 F1 分数。

系统	音频	Text	音频+文本
Poria 等人 [67]	60.1%	79.1%	80。1%
Zadeh 等人 [68]	67.4%	75.6%	-
Chen 等人 [69]	-	67.3%	-
Lian 等人 [44]	-	-	80.8%
HCAM（本文工作）	**67。4**%	**85。4**%	**85.8**%

表 VI：我们系统采用不同训练范式的加权F1分数

训练范式	IEMOCAP		MELD	CMU-MOSI
	四种方式	6 类		
层次化	85.9%	70.5%	65.8%	85.8%
非层次化	82.3%	68。5%	65%	84.4%

其中，我们实现了 8.5% 的相对提升。随后，经过融合模块后，我们又获得了约 4% 的相对提升。值得注意的是，在三个数据集中，MELD 是类别分布最不平衡的数据集，其单次对话中的说话人数量也最多。然而，与一些先前的研究工作 [63] 不同，我们的模型在情感建模过程中不依赖任何说话人信息。

在表 V 中，我们将所提出的模型与其他研究在 CMU-MOSI 数据集上的表现进行了比较。由于该数据集规模较小（仅包含 2199 个话语），所提出的模型在该数据集上容易过拟合。不过，我们注意到，该模型在音频模态上的表现与此前报告的仅使用音频输入所得的最佳结果相当。与此前报告的纯文本性能最佳结果相比，我们将性能相对提升了 30%；而同时使用音频与文本的模型，则较此前的最先进性能相对提升了 26%。

B. 分层建模的影响

为了了解我们分层建模方法的优势，我们调整了训练范式，将训练的第二阶段和第三阶段合并。训练范式调整后的结果如表 VI 所示。我们注意到，在全部四种数据集设置下，采用所提出的分层建模方法都能取得更好的性能，且模型架构无需改变。这些实验

表 VII：从上下文GRU中移除自注意力块后，系统在两种模态上的加权F1分数。

模态	第二阶段模型	IEMOCAP		MELD	CMU-MOSI
		四分类	6 类		
音频	不含自注意力	75.2%	63.8%	50.1%	67%
	双向门控循环单元与自注意力	78.7%	65.7%	50.1%	67.4%
Text	不含双向门控循环单元自注意力	79.9%	62.5%	65%	85.8%
	双向门控循环单元与自注意力	81.4%	64。4%	65.6%	85.4%

凸显了 HCAM 框架中所提出的各阶段采用课程式设计的优势。

C. 阶段*II*中自注意力的重要性

为了了解自注意力在上下文GRU中的作用，我们开展了一项消融实验，从上下文GRU中移除了自注意力块。该实验的结果如表VII所示。我们注意到，在IEMOCAP中，自注意力提升了两种模态的性能。这是意料之中的，因为IEMOCAP中的对话长度更长（对话长度为110个话语）。不同于其他数据集，在CMU-MOSI 数据集上引入自注意力块后，我们观察到文本模态的绝对性能下降了0.4%。这可能部分是由于该数据集规模较小，导致双向门控循环单元模型出现过拟合。

D. 使用自动语音识别生成的转录文本进行测试

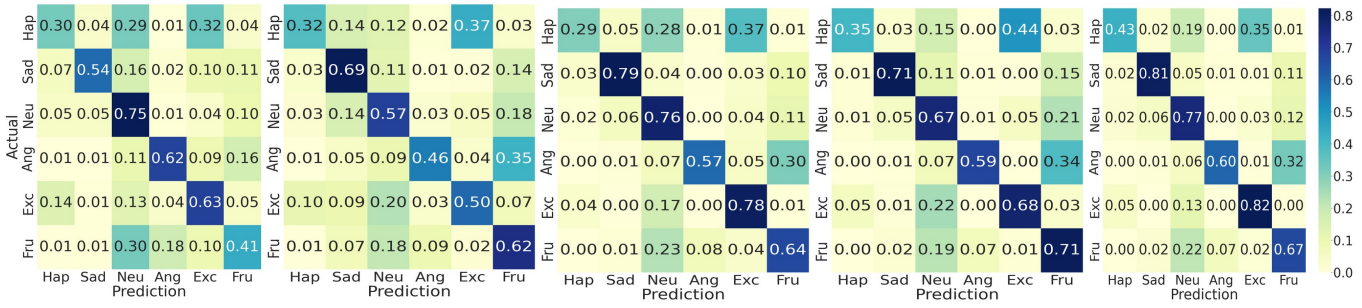
为了了解我们的模型对文本模态中噪声的鲁棒性，我们使用自动语音识别转录文本替代所提供的转录文本对模型进行测试。模型的训练过程与之前的实验相比没有改变

我们使用一个预训练的自动语音识别系统 [2] 来提供转录文本。表 VIII 报告了该自动语音识别系统在三个数据集上的词错误率（WER）。由此可见，情感对话语音上的词错误率显著高于其他受控数据集。MELD 的自动语音识别性能最低，部分原因是该数据集中存在较高水平的背景噪声。尽管词错误率较高，该模型在 IEMOCAP 和 CMU-MOSI 数据集上仍表现良好（如表IX所示，IEMOCAP 六分类的绝对性能下降为 2.4%，CMU-MOSI 则为 5.7%）。在此推理设置下的基线结果方面，由于我们尚未在其他地方找到类似的推理设置，因此采用了我们先前的工作 [47], 。与先前的工作相比，我们实现了 32% 的显著相对提升。

[2]https://cloud.google.com/speech-to-text

表 VIII：所用三个数据集上 Google 自动语音识别系统的词错误率（WER）（%）。所用缩写：快乐：*Hap.*，中性：*Neu.*，愤怒：*Ang.*，兴奋：*Exc.*，沮丧：*Fru.*，厌恶：*Dis.*，积极：*Pos.*和消极：*Neg.*

数据集	IEMOCAP						MELD							CMU-MOSI	
划分	Hap.	Sad	Neu.	Ang.	Exc.	Fru.	Ang.	Sad	Neu.	Fear	Sur.	Dis.	Joy	Pos.	Neg.
训练集	36	44	31.2	23.4	36.5	28	50.1	50.1	50.2	54.8	59	47.5	54.1	37.3	37.9
Val							47.4	49.9	46.4	52.8	58.8	46.6	50.7	34	37.9
Test	34.5	47.5	30.2	27.3	36	30.5	50.5	49.2	48.1	59.5	61.2	49.3	54.5	38.2	40.4



(a) 音频（阶段 1） (b) 文本（阶段 1） (c) 音频（阶段 2） (d) 文本（阶段 2） (e) 融合（阶段 3）

图4：我们的模型在包含6个类别的 IEMOCAP 数据集上运行时，不同阶段的混淆矩阵。所用缩写：快乐：快乐，中性：中性，愤怒：愤怒，兴奋：兴奋，沮丧：沮丧

表IX：测试时，我们的系统使用自动语音识别转录文本在各数据集上的加权F1分数。

文本转录 使用测试时间	IEMOCAP		MELD	CMU-MOSI
	四分类	六分类		
ASR	84.6%	68.1%	50.2%	80.1%
原始	85.9%	70。5%	65。8%	85.8%
此前最优水平 [47]	77.3%	-	-	-

表X：有无监督对比损失时的加权 F1 分数（%）。

Loss	IEMOCAP		MELD	CMU-MOSI
	四分类	六分类		
CE	84。5%	69.9%	65.5%	84.7%
交叉熵+	85.9%	70.5%	65。8%	85.8%
Supcon				

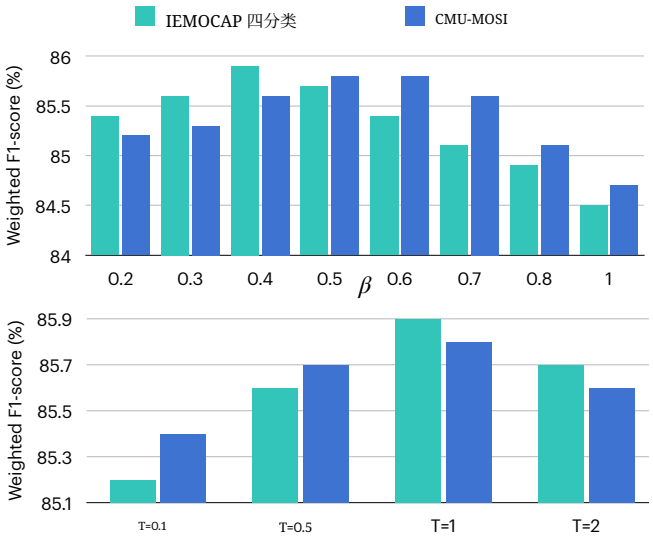


图 5：对于 IEMOCAP 四分类和 CMU-MOSI 数据集，测试性能随 β（公式 8）及监督式对比损失中的温度参数变化而产生的变化。

E. 监督式对比损失的作用

在不使用和使用监督式对比损失的情况下，对所提出模型的性能进行了评估。结果如表X所示。对于所有数据集，我们观察到，引入监督式损失后，模型的性能有所提升。

然而，监督式对比损失带来的性能提升因数据集而异。添加监督式对比损失后，IEMOCAP-6 和 MELD 的性能分别仅提升了 0.6% 和 0.3%（绝对值）。当类别数量减少时，引入监督式对比损失能够显著提升模型的性能（在 IEMOCAP 四分类中分别提升了 1.4% 和 1.1% 分别对应 CMU-MOSI）。该损失涉及两个超参数，即温度以及用于与交叉熵损失结合的权重。图 5 展示了在 IEMOCAP 四分类和 CMU-MOSI 数据集上，这些参数变化对测试性能的影响。

F. 不同情绪类别的性能

为了展示我们的模型识别不同类别的性能，我们在图 4 中展示了 IEMOCAP 六分类的混淆矩阵。可以看出，

表 XI：推理时集成与不进行推理时集成的加权 F1 分数

模态	推理时间集成	IEMOCAP 四分类	IEMOCAP 六分类	MELD	CMU-MOSI
音频（第二阶段）	Yes	78.7%	65。7%	50.1%	67.4%
	No	78.7%	65。7%	49.6%	63%
文本（第二阶段）	Yes	81.4%	64.4%	65.6%	85.4%
	No	81。4%	64.4%	65.4%	85.4%
音频 +文本 (第三阶段)	Yes	85.9%	70.5%	65.8%	85。8%
	No	85.3%	70.3%	65.5%	85.8%

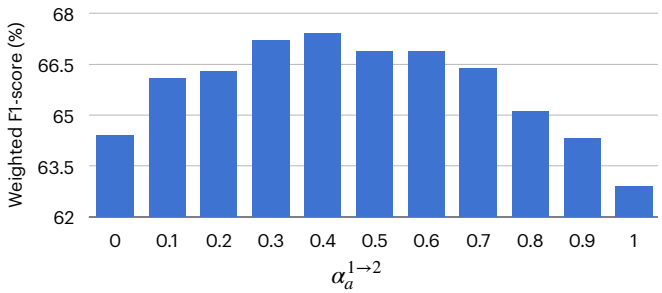


图 6：对于 CMU-MOSI，随着 $\alpha_a^{1\to2}$(Eq.9) 变化，测试性能的变化情况。

模型在区分快乐和兴奋类别时存在相当大的误差。这在一定程度上是意料之中的，因为这些情绪类别彼此密切相关。混淆矩阵还表明，所提出的模型不会偏向任何特定的情绪类别。

G. 测试时集成的作用

如第IV-E节所述，我们对模型在特定训练阶段的预测结果与前一阶段的预测结果进行加权组合。表 XI 展示了这种测试时组合策略的效果。我们将该策略与在推理的第二阶段和第三阶段不使用任何集成方法的情况进行了比较。我们发现，对于 CMU-MOSI，该策略使音频模态在第二阶段的性能提升了 4.4%。虽然 MELD 的性能略有提升，但集成预测结果并未使 IEMOCAP 的性能得到任何改善。对于第三阶段的组合，除 CMU-MOSI 外，我们发现其他所有测试设置的性能均略有提升。图 6 展示了 CMU-MOSI 在第二阶段结束时音频模态测试性能的变化情况。

七、结论

本文针对对话情感识别任务提出了一种分层多模态神经架构。所提出的架构首先学习话语级语音和文本的表示，然后通过对录音中的其他话语采用自注意力机制，进一步改进这些表示。利用协同注意力机制，实现两种模态的有效多模态融合。在三个基准在这些数据集上，我们取得了新的最先进的结果。我们进一步展示了模型在使用 ASR 生成的文本转录进行测试时的鲁棒性。通过广泛的消融研究，我们还展示了建模框架的不同方面以及超参数选择所产生的影响。未来，我们计划扩展这些方法，将视觉模态也纳入情感识别中。

参考文献

[1] Maja Pantic、Nicu Sebe、Jeffrey F Cohn 和 Thomas Huang，《情感多模态人机交互》，载于*ACM* 国际多媒体会议，2005年，页码 669–676。[2] Bharat Gaind、Varun Syal 和 Sneha Padgalwar，《社交媒体上的情绪检测与分析》，*arXiv* 预印本 *arXiv:1901.08458*，2019年。[3] Bryan Li、Dimitrios Dimitriadis 和 Andreas Stolcke，《客服通话中的声学与词汇情感分析》，载于*ICASSP*。IEEE，2019年，页码 5876–5880。[4] Surjya Ghosh、Sumit Sahu、Niloy Ganguly、Bivas Mitra 和 Pradipta De，《Emokey：一款用于心理健康监测的情绪感知智能手机键盘》，载于*COMSNETS*。IEEE，2019年，页码 496–499。[5] Soujanya Poria、Navonil Majumder、Rada Mihalcea 和 Eduard Hovy，《对话中的情绪识别：研究挑战、数据集与最新进展》，*IEEE Access*，第 7 卷，页码 100943–100953，2019年。[6]Paweł Tarnowski、Marcin Kołodziej、Andrzej Majkowski 和 Remigiusz J Rak，《利用面部表情进行情绪识别》，*ProcediaComputerScience*，第 108 卷，页码 1175–1184，2017年。[7] Klaus R Scherer、Tom Johnstone 和 Gundrun Klasmeyer，情绪的声音表达。牛津大学出版社，2003年。[8] Costanza Navarretta，《交际性身体行为中的个体性》，载于认知行为系统，页码 417–423。施普林格，2012年。[9] R Benjamin Knapp、Jonghwa Kim 和 Elisabeth Andr´e，《生理信号及其在人机交互中增强情绪识别方面的应用》，载于面向情绪的系统，页码 133–159。施普林格，2011年。[10] Soujanya Poria、Erik Cambria、Rajiv Bajpai 和 Amir Hussain，《情感计算综述：从单模态分析到多模态融合》，*InformationFusion*，第 37 卷，页码 98–125，2017年。[11] Yelin Kim 和 Emily Mower Provost，《Isla：音视频情绪识别的时间分段与标注》，*IEEE Transactions on Affective Computing*，第 10 卷，第 2 期，页码 196–208，2017年。[12] Hengshun Zhou、Jun Du、Yuanyuan Zhang、Qing Wang、Qing-Feng Liu 和 Chin-Hui Lee，《注意力网络中的信息融合：采用自适应多层因子化双线性池化进行音视频情绪识别》，*IEEE/ACM Transactions on Audio, Speech, and LanguageProcessing*，第 29 卷，页码 2617–2629，2021年。[13] Seunghyun Yoon、Seokhyun Byun 和 Kyomin Jung，《利用音频与文本进行多模态语音情绪识别》，载于*SLT*。IEEE，2018年，页码 112–118。[14] Leonardo Pepino、Pablo Riera、Luciana Ferrer 和 Agust´ın Gravano，《利用声学与文本特征进行语音情绪识别的融合方法》，载于*ICASSP*。IEEE，2020年，页码 6484–6488。[15] Trisha Mittal、Uttaran Bhattacharya、Rohan Chandra、Aniket Bera 和 Dinesh Manocha，《M3er：利用面部、文本和语音线索进行乘性多模态情绪识别》，载于*AAAI*，2020年，第 34 卷，页码 1359–1367。[16] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli，《wav2vec 2.0：用于语音表征自监督学习的框架》，神经信息处理系统进展，第 33 卷，页码 12449–12460，2020年。[17] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov，《RoBERTa：一种鲁棒优化的 BERT 预训练方法》，*arXiv* 预印本 *arXiv:1907.11692*，2019年。[18] Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan，《有监督对比学习》，神经信息处理系统进展，第 33 卷，页码 18661–18673，2020年。[19] 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》，*LREC*，第 42 卷，第 4 期，页码 335–359，2008年。

[20] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea，《MELD：用于对话情绪识别的多模态多方数据集》，载于 ACL，2019年，页码 527–536。[21] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency，《MOSI：在线观点视频中情感强度与主观性分析的多模态语料库》，arXiv 预印本 arXiv:1606 06259，2016年。。[22] Disa A Sauter、Frank Eisner、Andrew J Calder 和 Sophie K Scott，《非言语声音情绪表达中的感知线索》，《实验心理学季刊》，第 63 卷，页码 2272，2010年 11，页码 2251。[23] Iker Luengo、Eva Navas、Inmaculada Hern´aez 和 Jon S´anchez，《利用韵律参数进行自动情绪识别》，载于 Interspeech，2005年。[24] Shashidhar G Koolagudi、Nitin Kumar 和 K Sreenivasa Rao，《利用音段级韵律分析进行语音情绪识别》，载于 ICDeCom。IEEE，2011年，页码 1–5。[25] Florian Eyben、Felix Weninger 和 Bj¨orn Schuller，《真实环境声学条件下的情感识别——特征选择的新视角》，载于 Interspeech，2013年。[26] Bj¨orn Schuller、Stefan Steidl、Anton Batliner、Alessandro Vinciarelli、Klaus Scherer、Fabien Ringeval、Mohamed Chetouani、Felix Weninger、Florian Eyben、Erik Marchi 等，《Interspeech 2013 计算副语言学挑战赛：社会信号、冲突、情绪、自闭症》，载于 Interspeech，2013年。[27] Florian Eyben、Klaus R Scherer、Bj¨orn W Schuller、Johan Sundberg、Elisabeth Andr´e、Carlos Busso、Laurence Y Devillers、Julien Epps、Petri Laukka、Shrikanth S Narayanan 等，《用于语音研究与情感计算的日内瓦极简声学参数集（GeMAPS）》，IEEE Transactions on Affective Computing，第 7 卷，第 2 期，页码 190–202，2015年。[28] Florian Eyben、Martin W¨ollmer 和 Bj¨orn Schuller，《openSMILE：慕尼黑多功能快速开源音频特征提取器》，载于第 18 届 ACM 国际多媒体会议论文集，2010年，页码 1459–1462。[29] Mirco Ravanelli 和 Yoshua Bengio，《利用 sincnet 从原始波形进行说话人识别》，载于 SLT。IEEE，2018年，页码 1021–1028。[30] Neil Zeghidour、Olivier Teboul、F´elix de Chaumont Quitry 和 Marco Tagliasacchi，《LEAF：用于音频分类的可学习前端》，ICLR，2021年。[31] Purvi Agrawal 和 Sriram Ganapathy，《利用深度变分网络进行调制滤波器学习，以实现鲁棒语音识别》，《IEEE信号处理精选主题期刊》，第 13 卷，第 2 期，页码 244–253，2019年。[32] Purvi Agrawal 和 Sriram Ganapathy，《基于相关性加权的语音和音频信号可解释表示学习》，IEEE/ACM Transactions on Audio, Speech, and Language Processing，第 28 卷，页码 2823–2836，2020年。[33] Steffen Schneider、Alexei Baevski、Ronan Collobert 和 Michael Auli，《wav2vec：用于语音识别的无监督预训练》，Interspeech，页码 3465–3469，2019年。[34] Shamane Siriwardhana、Andrew Reis、Rivindu Weerasekera 和 Suranga Nanayakkara，《联合微调“类BERT”自监督模型以提升多模态语音情绪识别》，Interspeech，页码 3755–3759，2020年。[35] Manon Macary、Marie Tahon、Yannick Est`eve 和 Anthony Rousseau，《论自监督预训练声学与语言特征在连续语音情绪识别中的应用》，载于 SLT。IEEE，2021年，页码 373–380。[36] Leonardo Pepino、Pablo Riera 和 Luciana Ferrer，《利用 wav2vec 2.0 嵌入进行语音情绪识别》，Interspeech，页码 3400–3404，2021年。[37] Rui Xia 和 Chengqing Zong，《探索词语关系特征在情感分类中的应用》，载于 Coling 2010：海报论文，2010年，页码 1336–1344。[38] Tom´aˇs Mikolov、Edouard Grave、Piotr Bojanowski、Christian Puhrsch 和 Armand Joulin，《分布式词表示预训练的进展》，载于 LREC，2018年。[39] Navonil Majumder、Devamanyu Hazarika、Alexander Gelbukh、Erik Cambria 和 Soujanya Poria，《利用上下文建模和层次融合进行多模态情感分析》，《知识型系统》，第 161 卷，页码 124–133，2018年。[40] Yash Sharma、Gaurav Agrawal、Pooja Jain 和 Tapan Kumar，《利用 GloVe 进行情感分析的词向量表示》，载于 ICCT。IEEE，2017年，页码 279–284。[41] Haixia Liu，《利用 word2vec 对引文进行情感分析》，arXiv 预印本 arXiv:1704.00177，2017年。

[42] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，《Attention Is All You Need》，《神经信息处理系统进展》，第30卷，2017年。[43] Jacob Devlin、Ming Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：面向语言理解的深度双向 Transformer 预训练》，载于 NAACL-HLT，2019年，第4171–4186页。[44] Zheng Lian、Bin Liu 和 Jianhua Tao，《SMIN：用于对话情感识别的半监督多模态交互网络》，IEEE Transactions on Affective Computing，2022年。[45] Soujanya Poria、Erik Cambria 等，《用户生成视频中的上下文相关情感分析》，载于 ACL（第1卷：长篇论文），2017年，第873–883页。[46] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J Zico Kolter、Louis-Philippe Morency 和 Ruslan Salakhutdinov，《用于未对齐多模态语言序列的多模态 Transformer》，载于 ACL。美国国立卫生研究院公共访问，2019年，第2019卷，第6558页。[47] Soumya Dutta 和 Sriram Ganapathy，《结合可学习前端和自注意力的情绪识别多模态 Transformer》，载于 ICASSP。IEEE，2022年，第6917–6921页。[48] Amir Zadeh、Paul Pu Liang、Navonil Mazumder、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency，《面向多视图序列学习的记忆融合网络》，载于 AAAI，2018年，第32卷。[49] AmirAli Bagher Zadeh、Paul Pu Liang、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency，《开放环境下的多模态语言分析：CMU-MOSEI数据集与可解释的动态融合图》，载于 ACL（第1卷：长篇论文），2018年，第2236–2246页。[50] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee，《ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表示》，《神经信息处理系统进展》，第32卷，2019年。[51] Hao Tan 和 Mohit Bansal，《LXMERT：从 Transformer 中学习跨模态编码器表示》，载于 EMNLP-IJCNLP，2019年，第5100–5111页。[52] Navonil Majumder、Soujanya Poria、Devamanyu Hazarika、Rada Mihalcea、Alexander Gelbukh 和 Erik Cambria，《DialogueRNN：一种用于对话中情感检测的注意力循环神经网络》，载于 AAAI，2019年，第33卷，第6818–6825页。[53] Deepanway Ghosal、Navonil Majumder、Soujanya Poria、Niyati Chhaya 和 Alexander Gelbukh，《DialogueGCN：一种用于对话中情绪识别的图卷积神经网络》，载于 EMNLP-IJCNLP，2019年，第154–164页。[54] Yoshua Bengio、J´erˆome Louradour、Ronan Collobert 和 Jason Weston，《课程学习》，载于《第26届机器学习国际年会论文集》，2009年，第41–48页。[55] Jacob Kahn、Morgane Riviere、Weiyi Zheng、Evgeny Kharitonov、Qiantong Xu、Pierre-Emmanuel Mazar´e、Julien Karadayi、Vitaliy Liptchinsky、Ronan Collobert、Christian Fuegen 等，《Libri-Light：一种用于有限或无监督自动语音识别的基准》，载于 ICASSP。IEEE，2020年，第7669–7673页。[56] Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis Tyers 和 Gregor Weber，《Common Voice：一个大规模多语言语音语料库》，载于 LREC，2020年，第4218–4222页。[57] John J Godfrey、Edward C Holliman 和 Jane McDaniel，《Switchboard：用于研究与开发的电话语音语料库》，载于 ICASSP。IEEE，1992年，第517–520页。[58] Christopher Cieri、David Miller 和 Kevin Walker，《Fisher语料库：面向下一代语音转文本技术的资源》，载于 LREC，2004年，第4卷，第69–71页。[59] Sijie Mai、Haifeng Hu 和 Songlong Xing，《分而治之，再行融合：一种面向多模态情感计算、兼顾局部与全局视角的层次特征融合网络》，载于 ACL，2019年，第481–492页。[60] Runnan Li、Zhiyong Wu、Jia Jia、Yaohua Bu、Sheng Zhao 和 Helen Meng，《迈向语音情绪识别的判别式表示学习》，载于 IJCAI，2019年，第5060–5066页。[61] Sijie Mai、Songlong Xing 和 Haifeng Hu，《一种面向多模态人类情感计算、具有全局视角的局部受限模态融合网络》，IEEE多媒体汇刊，第22卷，第1期，第122–137页，2019年。[62] Weizhou Shen、Junqing Chen、Xiaojun Quan 和 Zhixian Xie，《DialogXL：用于多方对话情绪识别的一体化 XLNet》，载于 AAAI，2021年，第35卷，第13789–13797页。[63] Yuzhao Mao、Guang Liu、Xiaojie Wang、Weiguo Gao 和 Xuan Li，《DialogueTRM：探索对话中的多模态情感动态。

”于*EMNLP*，2021，第 2694–2704 页。

[64] Zaijing Li、Fengxiao Tang、Ming Zhao 和 Yusen Zhu，《EmoCaps：一种基于情感胶囊的对话情感识别模型》，载于计算语言学协会：*ACL 2022*年论文集，爱尔兰都柏林，2022年5月，第1610–1618页，计算语言学协会。[65] Dong Zhang、Liangqing Wu、Changlong Sun、Shoushan Li、Qiaoming Zhu 和 Guodong Zhou，《对多说话者对话中的情绪检测同时建模上下文和说话者敏感依赖关系》，载于*IJCAI*，2019年，第5415–5421页。[66] Jingye Li、Donghong Ji、Fei Li、Meishan Zhang 和 Yijiang Liu，《Hitrans：一种基于 Transformer 的上下文和说话者敏感模型，用于对话中的情绪检测》，载于第*28*届国际计算语言学会议论文集，2020年，第4190–4200页。[67] Soujanya Poria、Erik Cambria、Devamanyu Hazarika、Navonil Mazumder、Amir Zadeh 和 Louis-Philippe Morency，《用于上下文多模态情感分析的多层多重注意力机制》，载于*ICDM*。IEEE，2017年，第1033–1038页。[68] Amir Zadeh、Minghai Chen、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency，《用于多模态情感分析的张量融合网络》，载于*EMNLP*，2017年，第1103–1114页。[69] Minghai Chen、Sen Wang、Paul Pu Liang、Tadas Baltruˇsaitis、Amir Zadeh 和 Louis-Philippe Morency，《结合词级融合与强化学习的多模态情感分析》，载于*ICMI*，2017年，第163–171页。