基于情感信息辅助的多模态情绪识别

吴良庆　刘启元　张栋†　王建成　李寿山　周国栋

苏州大学计算机科学与技术学院，江苏 苏州 215006；

† 通讯作者，E-mail: dzhang17@stu.suda.edu.cn

摘要　近年来，多模态情感分析成为自然语言处理的热点研究领域，挖掘多模态内容（如视频和语音等）包含的情绪或情感信息具有十分重要的现实意义。基于多模态特征的情绪分类和情感分类作为情感分析的两个子任务，已有大量工作对两者进行单独研究，但是在多模态领域，还没有相关研究利用情感信息帮助识别说话人的情绪。不同于纯文本的情绪分析，本文面向多模态数据（文本和语音）进行情绪识别研究。为了同时考虑多模态数据特征，我们提出一种新颖的联合学习框架，将多模态情绪分类作为主任务，多模态情感分类作为辅助任务，通过情感信息来辅助提升情绪识别任务的性能。具体而言，通过私有网络层对主任务中的文本和语音模态信息分别进行编码，以学习单个模态内部的情绪独立特征表示。接着，通过辅助任务中的共享网络层来获取主任务的辅助情绪表示以及辅助任务的单模态完整情感表示。在得到主任务的文本和语音辅助情绪表示之后，分别与主任务中的单模态独立特征表示进行结合，得到主任务中单模态情绪信息的完整表示。最后通过自注意力机制捕捉每个任务上的多模态交互特征，得到最终的多模态情绪表示和情感表示。实验结果表明，我们的方法在多模态情感分析数据集上可以通过情感辅助信息大幅度提升情绪分类任务的性能，同时情感分类任务的性能也得到了一定的提升。

关键词　多模态；情绪识别；联合学习；自然语言处理

中图分类号　TP391

Multimodal Emotion Recognition with Auxiliary Sentiment Information

WU Liangqing，LIU Qiyuan，ZHANG Dong†，WANG Jiancheng，LI Shoushan，ZHOU Guodong

School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006;

†Corresponding Author, E-mail: dzhang17@stu.suda.edu.cn

Abstract　In recent years, multimodal sentiment analysis has been a hot research area in natural language processing. Mining the emotion or sentiment information inside multimodal contents has a great impact on real-life applications. As two subtasks in sentiment analysis, multimodal emotion and sentiment classification have been studied separately. However, there's no existing work on exploring sentiment information to help detect the emotion expressed by the speaker in multimodal area. Different from the previous studies with only text, this paper focuses on multimodal data (text and audio) to perform emotion recognition. To simultaneously address the characteristics of multimodal data, we propose a novel joint learning framework, which allows auxiliary task (multimodal sentiment classification) to help the main task (multimodal emotion classification).Specifically, private neural layers are designed for text and audio modalities from the main task to learn the uni-modal independent dynamics. Secondly, with the shared neural layers from auxiliary task, we obtain the uni-modal representations of the auxiliary task and the auxiliary representations of the main task. We combine the uni-modal independent dynamics with the auxiliary representations for each modality to acquire the uni-modal representations of the main task. Finally, in order to capture multi-modal interactive dynamics, we fuse the text and audio modalities' representations for the main and auxiliary tasks separately to obtain the final multimodal emotion and sentiment representations with the self attention mechanism. Empirical results

国家自然科学基金(61331011， 61375073)资助

收稿日期：0000-00-00；修回日期：0000-00-00；网络出版时间：

网络出版地址：

demonstrate the effectiveness of our approach to multimodal emotion classification task as well as the sentiment classification task.

Key words Multi-modal; Emotion recognition; Joint learning; Natural language processing

随着头条、抖音等社交媒体平台的迅速崛起，人们越来越喜欢上传语音或视频等多模态内容来表达自己的情感（Sentiment）和情绪（Emotion）。挖掘多模态内容所包含的情感信息对于舆情发现和用户反馈等应用具有十分重要的意义[1][2]。因此，越来越多的研究者专注于分析多模态数据的情感倾向，而多模态的情绪分析作为细粒度的情感分析备受关注。

值得注意的是，在纯文本情绪分析领域，已有较多研究工作利用情感信息帮助识别文本的情绪标签，同时在一定程度上也可以提升情感的分类性能[3]-[4]。这些工作认为句子的情感标签和情绪标签存在依赖关系，例如："*Oh umm, my big scene is coming up. Big scene coming up.*"，本例的情绪标签为喜悦（*Joy*），情感标签为正向（*Positive*）。例子中的场景是说话人在兴奋地等待他所演的镜头在荧幕上出现的一刻，如果我们判定句子所表达的情绪为喜悦，那么可以容易识别本例中的正向情感。反之，如果我们判定句子的情感是正向，则可以淘汰掉一些负面的情绪，如愤怒（*Anger*）、厌恶（*Disgust*）和悲伤（*Sadness*）等。因此，采用联合学习方法可以有效地捕捉情感和情绪的共享信息，达到情感、情绪分类任务同时提升的目的。然而，在多模态情感分析领域，目前还没有相关工作利用情感和情绪分类任务进行联合学习。

不同于纯文本分析，如何利用多模态特征针对情感与情绪任务进行联合学习是一大挑战[5]。一方面，不同的模态存在各自的特性，我们称之为单模态独立特征（Uni-modal independent dynamics）；另一方面，两个模态存在交互关系，我们称之为多模态交互特征（Multi-modal interactive dynamics）。例如："*I got no sleep last night!*"，我们仅通过文本信息很难判断其准确的情绪，必须同时考虑其对应的语音信息，若语音信息表现出大声的特征，则可以确定该例子的情绪为愤怒，若语音信息表现出低沉的语气，则可以认为该例子包含悲伤的情绪。因此，我们需要准确地捕捉文本和语音之间的联系，更好地对单模态独立特征和多模态交互特征进行建模。

为了解决以上的挑战，本文面向文本和语音模态数据，提出一种基于情感信息辅助的多模态情绪识别方法，该方法通过联合学习的方式可以同时提升情感分类和情绪识别任务的性能。具体而言：1）首先，将情绪分类作为主任务，其文本和语音信息分别经过私有 LSTM 网络，从而学习单个模态内部的情绪独立特征表示。2）其次，将情感分类作为辅助任务，通过辅助任务中的共享 LSTM 层来获取主任务的辅助情绪表示以及辅助任务的单模态完整情感表示。3）在得到主任务的文本和语音辅助情绪表示之后，分别与主任务中的单模态独立特征表示进行结合，得到主任务中单模态情绪信息的完整表示，从而增强主任务的情绪特征表示。4）在此基础上，为了捕捉每个任务上多模态的交互特征，每个任务均利用自注意力机制加权融合来自多模态数据的表示。最终，每个任务根据各自的多模态加权融合表示进行对应的情绪/情感分类。实验结果表明，我们的方法可以在多模态情感分析数据集上同时提升情绪分类和情感分类任务的性能。

本文第 2 节介绍联合学习以及多模态情感分析的相关工作；第 3 节详细描述本文提出的基于联合学习框架的多模态情感分析方法；第 4 节介绍实验设置以及实验的结果和分析；第 5 节给出结语。

1 相关工作

1.1 情绪与情感分析（单文本）

近年来，不少研究工作采用联合学习框架在情感分析上取得成效。在属性级情感分析任务上，Ma 等[6]通过将属性抽取和属性级情感分类两个任务进行联合学习极大提升了属性级情感分类任务的性能。Chen 等[7]提出联合模型针对情绪分类和情绪原因识别两个任务同时进行建模，从而有效提升了两个任务的识别性能。

情绪分类和情感分类是情感分析中两个不同的子任务，由于情绪标签和情感标签有着强烈的联系，因此两个任务是紧密相关的。Gao 等[3]通过标注一个额外的带有情绪和情感标签的数据集来提升两个任务的性能。然而，在现实场景下难以获取类似的数据集。Wang 等[4]采用整数线性规划（ILP）对情绪和情感分类任务进行联合学习，通过约束条件获取情绪分类器的输出和情感分类器的输出之间的联系。

不同于以往的联合学习研究，在本文我们将联合学习框架扩展到对包含文本和语音信息的多模态内容

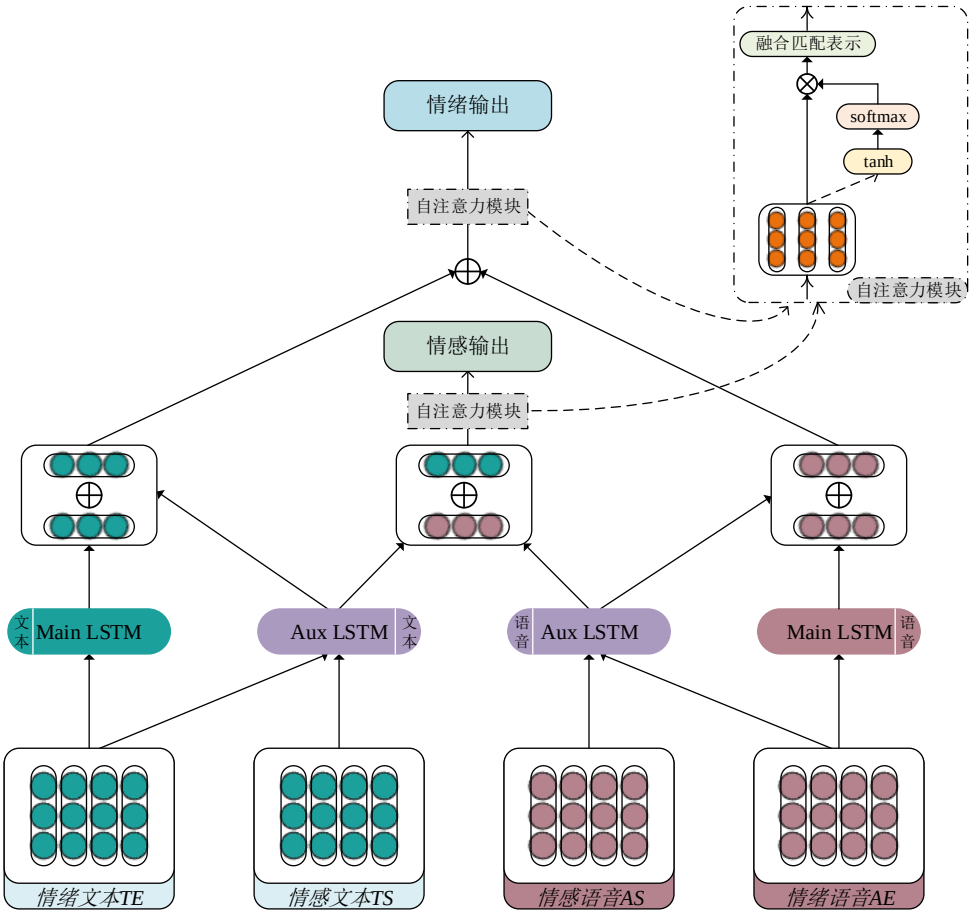

图 1　联合多模态情绪与情感学习网络的模型框架图

Fig.1 Overall architecture of joint multimodal emotion and sentiment learning network

进行探索。

1.2 情绪与情感分析（多模态）

多模态情感分析结合语言以及非语言的信息去探测用户所表达的情感，现已成为了一个热点研究课题。早期的方法为了简化问题采用简化时序信息的操作对多模态内容进行情感分析[8]，Morency 等[1]通过将时序信息进行平均得到每个模态信息的表示后进行拼接，并把得到的多模态信息作为隐马尔科夫模型（HMMs）的输入特征，获取最终的情感信息。类似地，Zadeh 等[2]和 Pérez-Rosas 等[9]采用拼接的多模态特征训练支持向量机（SVM）模型，然而平均的信息无法有效捕捉模态内部的信息和模态之间的交互信息，因此这些方法难以对多模态信息进行有效建模。

最近的多模态情感分析研究工作主要采用深度学习模型对模态内部信息和模态之间的交互信息进行建模。Zadeh 等[5]采用张量融合网络的方法，通过创建一个多维张量来捕捉多模态内容所包含的情感信息。Liu 等[10]在张量融合网络的基础上提出低阶的张量网络，在提升性能的同时减少了原模型的参数以及计算复杂度。Bertero 等[11]对长短期记忆网络（LSTM）进行扩展，对多模态内容进行情绪识别。Zadeh 等[12]提出图记忆融合网络，通过门控记忆单元存储模态内部信息以及模态之间的交互信息，并加入动态融合图来反映有效的情绪信息。Hazarika 等[13]采用多层门控循环单元（GRU）来存储当前多模态内容的说话人信息以及上下文信息，并把 GRU 的输出作为全局信息来分析当前多模态内容所包含的情绪信息。

虽然以上的方法取得了不错的效果，但是这些方法都只考虑了多模态内容中的情绪信息或者情感信息，却忽略了情绪与情感之间的联系。因此，在本文我们提出一种基于情感信息辅助的多模态情绪识别方法，该方法通过联合学习的方式可以同时提升情感分类和情绪识别任务的性能。

2　方法

本节中，我们将详细介绍基于情感信息辅助的多模态情绪分类方法，该方法通过联合学习的方式来融合多模态数据集上的情绪和情感信息，将情绪分类作为主任务，情感分类作为辅助任务，通过情感分类任务来辅助情绪分类任务。图 1 为本文所提出的基于联合学习方法的模型框架图，该模型包含以下几个部分：1）特征表示层：将多模态内容的文本和语音信息映射成特征，作为神经网络的输入。2）任务私有/共享层：通过私有网络层对模态内部信息进行编码，通过共享网络层对主任务和辅助任务进行联合学习。3）模态交互层：对主任务的文本和语音特征进行融合，获取模态之间的交互信息，对辅助任务亦如此。4）预测分类层：使用单模态信息和多模态融合信息对主任务和辅助任务进行预测。

2.1 特征表示层

本节中，我们将介绍多模态情绪分类任务和情感分类任务所使用的文本和语音特征。

对于文本模态，假设一句话语由 N 个词组成，通过 GloVe 来获取对应的预训练词向量[14]，未知词通过随机向量来初始化，即每个词可表示为 $w \in \mathbb{R}^{d_t}$ 。对于辅助任务，我们采用双层卷积神经网络对词向量进行卷积，提取对应的文本特征 TS 。对于主任务，对词向量不做额外的预处理，使用 TE 表示。

对于语音模态，假设一句话对应的语音有 M 帧信息，通过 OpenSMILE[15]语音分析框架获取辅助任务的语音特征 AS 。对于主任务，我们采用基于梅尔频谱系数的语音特征，把每一帧信息映射成包含静态数据（static）、一阶导数（delta）、二阶导数（double delta）的 MFSC 系数特征[16]，记为 AE 。

2.2 任务私有/共享层

本节中，我们以文本模态的特征为例来描述任务私有/共享层。

首先，我们采用私有的长短期记忆网络（LSTM）对主任务的文本特征 TE 进行编码，具体表示如下：

$$h_{main}^{TE-i} = LSTM_{main}(w^{TE-i}) \tag{1}$$

其中，w^{TE-i} 是主任务中第 i 个词的词向量，$H_{main}^{TE} = [h_{main}^{TE-1}, h_{main}^{TE-2}, \cdots, h_{main}^{TE-N}] \in \mathbb{R}^{N \times d_{mem}}$ 是一句话语经过私有 LSTM 后的文本模态独立特征表示序列。类似地，我们可以得到主任务中语音模态独立特征表示序列 $H_{main}^{AE} \in \mathbb{R}^{N \times d_{mem}}$ 。

然后，我们采用共享的 LSTM 层对主任务和辅助任务的文本特征共同进行编码，具体表示如下：

$$h_{aux}^{TS-i} = LSTM_{aux}(w^{TS-i}) \tag{2}$$

$$h_{aux}^{TE-i} = LSTM_{aux}(w^{TE-i}) \tag{3}$$

其中，h_{aux}^{TS-i} 是辅助任务中第 i 个词经过共享层后的隐层编码，h_{aux}^{TE-i} 是主任务中第 i 个词经过共享层后的辅助编码。对应地，可以得到辅助任务的完整编码序列 $H_{aux}^{TS} \in \mathbb{R}^{N \times d_{mem}}$ 和主任务的辅助编码序列 $H_{aux}^{TE} \in \mathbb{R}^{N \times d_{mem}}$ 。我们对主任务和辅助任务的语音特征进行类似操作，得到辅助任务语音模态的完整编码序列 H_{aux}^{AS} 和主任务语音模态的辅助编码序列 H_{aux}^{AE} 。

2.3 模态交互层

在这一层中，我们先对联合学习层获取的 6 个编码序列进行最大池化操作（maxpooling，记为 mp）。以主任务中的文本模态为例，具体表示如下：

$$C_{main}^{TE} = mp(H_{main}^{TE}) \tag{4}$$

$$C_{aux}^{TE} = mp(H_{aux}^{TE}) \tag{5}$$

$$C^{TE} = concat(C_{main}^{TE}, C_{aux}^{TE}) \tag{6}$$

其中，$C_{main}^{TE} \in \mathbb{R}^{d_{mem}}$ 是主任务的文本模态独立特征表示序列经过最大池化后的编码表示，$C_{aux}^{TE} \in \mathbb{R}^{d_{mem}}$ 是主任务的文本辅助编码序列经过最大池化后的编码表示，$C^{TE} \in \mathbb{R}^{2 \times d_{mem}}$ 是主任务的文本模态的完整表示。类似地，

我们可以得到主任务的语音模态的完整表示 C^{AE}，接着把主任务的文本和语音完整表示进行拼接（*concat*），得到主任务的多模态融合表示 $C^{E} \in \mathbb{R}^{4\times d_{mem}}$。对于辅助任务，在得到文本和语音模态的最大池化表示 C^{TS} 和 C^{AS} 后，我们对其进行拼接操作，得到辅助任务的多模态融合表示 $C^{S} \in \mathbb{R}^{2\times d_{mem}}$。

接着，我们采用自注意力机制加权主任务的多模态融合表示，具体表示如下：

$$A^{E} = \text{softmax}(W_p \cdot (\tanh(W_q \cdot C^{E}))) \tag{7}$$

$$F^{E} = A^{E} \cdot C^{E} \tag{8}$$

$$R^{E} = mp(\tanh(W_r \cdot F^{E} + b_r)) \tag{9}$$

其中，A^{E} 代表主任务多模态融合表示中文本和语音信息的重要程度，通过 A^{E} 可以计算得到多模态加权融合信息的表示 F^{E}，利用最大池化操作获取最终的多模态情绪加权融合表示 $R^{E} \in \mathbb{R}^{d_{mem}}$，$W_p$、$W_q$ 和 W_r 是网络层的权重，b_r 是网络层的偏置。类似地，我们可以得到辅助任务最终的多模态情感加权融合表示 R^{S}。

表 1　MELD 中各情绪类别的样本数量分布

Table 1　Distributions of emotion categories in MELD

情绪	愤怒	厌恶	恐惧	喜悦	中性	悲伤	惊讶
训练集	1109	271	268	1743	4710	683	1205
验证集	153	22	40	163	470	111	150
测试集	345	68	50	402	1256	208	281

表 2　MELD 中各情感类别的样本数量分布

Table 2　Distributions of sentiment categories in MELD

情感	正向	负向	中性
训练集	2334	2945	4710
验证集	233	406	470
测试集	521	833	1256

2.4 预测分类层

对于主任务和辅助任务，分别采用各自的单模态完整表示和多模态加权融合表示对多模态数据进行情绪或情感的预测。

以多模态情绪分类（主任务）为例，通过 2.3 节我们得到了多模态的完整表示 R^{E}，我们通过 softmax 层对情绪类别进行预测，具体表示如下：

$$\hat{y}^{E} = \text{softmax}(W \cdot R^{E} + b) \tag{10}$$

其中，$\hat{y}^{E}$ 是文本和语音模态融合后得到的多模态情绪分类结果，W 和 b 是 softmax 层的权重与偏置。类似地，可以得到辅助任务的多模态情绪分类结果 $\hat{y}^{S}$，以及主辅任务的单模态文本和语音的分类效果 $\hat{y}^{TE}$、$\hat{y}^{AE}$、$\hat{y}^{TS}$ 和 $\hat{y}^{AS}$。

2.5 优化策略

在网络训练过程中，对主任务和辅助任务均采用交叉熵误差作为损失函数，具体表示如下：

$$Loss(\hat{y}, y) = -\sum_{k=1}^{K}\sum_{c=1}^{C} y_k^c \cdot \log \hat{y}_k^c \tag{11}$$

其中，K 是训练样本的总数，C 是类别的数目，$\hat{y}_k^c$ 是第 k 个样本的预测标签，y_k^c 是第 k 个样本的真实标签。对于不同的任务，$\hat{y} \in \{\hat{y}^{TE}, \hat{y}^{AE}, \hat{y}^{E}, \hat{y}^{TS}, \hat{y}^{AS}, \hat{y}^{S}\}$，我们对公式（11）产生的所有 *Loss* 进行线性组合作为

最终的优化目标函数，实验中采用 Adam 优化器[17]来优化模型的参数。

3 实验

本节中，我们将分析本文提出的联合学习方法在多模态情绪分类和情感分类任务上的效果。

3.1 实验设置

表 3 **MELD 情感分类实验结果（F1:%）**

Table 3 Results on sentiment classification in MELD

	文本模态				语音模态				多模态			
情感	正向	负向	中性	*w*-avg	正向	负向	中性	*w*-avg	正向	负向	中性	*w*-avg
TFN	50.50	52.77	75.71	63.36	24.28	29.27	**64.17**	45.07	54.96	57.18	73.34	64.51
MFN	52.52	53.67	71.79	62.16	20.62	43.74	57.38	45.69	55.07	54.76	75.84	64.97
BC-LSTM	56.28	57.91	74.23	65.44	18.86	45.74	60.09	47.28	57.49	61.02	73.84	66.49
CMN	52.10	57.86	74.73	64.83	17.34	40.91	63.08	46.88	55.14	58.12	75.27	65.78
ICON	54.15	56.93	75.25	65.19	16.76	45.20	61.64	47.44	56.60	59.37	76.46	67.04
Ours	**56.41**	**58.07**	**76.04**	**66.38**	**26.49**	**49.00**	58.04	**48.86**	**58.14**	**61.46**	**77.52**	**68.53**

表 4 **MELD 情绪分类实验结果（F1:%）**

Table 4 Results on emotion classification in MELD

文本模态								
情绪	愤怒	厌恶	恐惧	喜悦	中性	悲伤	惊讶	*w*-avg
TFN	**39.79**	7.79	0.00	50.72	70.8	25.19	49.20	54.65
MFN	32.40	9.90	3.33	50.00	73.75	18.47	47.98	54.43
BC-LSTM	37.76	9.76	2.27	50.67	74.39	24.32	40.00	55.14
CMN	38.88	4.60	3.88	46.36	74.21	**26.17**	47.92	55.43
ICON	32.25	15.84	2.70	50.35	73.96	24.23	44.71	54.82
Ours	28.69	**20.00**	**11.94**	**53.57**	**77.13**	14.11	**50.18**	**56.43**
语音模态								
TFN	**33.95**	0.00	0.00	5.69	61.23	2.78	25.89	37.84
MFN	33.23	0.00	0.00	4.15	63.88	0.96	23.26	38.35
BC-LSTM	15.00	0.00	0.00	12.92	65.30	3.67	**27.52**	38.65
CMN	33.45	0.00	0.00	8.99	66.32	0.96	14.04	39.31
ICON	18.74	0.00	0.00	21.96	**66.38**	1.90	19.28	40.03
Ours	31.06	**4.84**	0.00	**22.57**	64.13	**4.48**	22.22	**41.32**
多模态								
TFN	42.33	8.06	2.63	47.84	74.65	20.51	46.84	55.82
MFN	36.99	7.41	3.23	51.00	74.90	17.82	46.48	55.47
BC-LSTM	40.58	2.78	2.13	46.12	74.35	**27.47**	**50.46**	55.98
CMN	42.29	2.82	3.6	49.46	75.15	25.00	50.00	56.89
ICON	41.72	6.74	5.56	50.72	74.90	24.93	49.90	57.01
Ours	**45.17**	**26.36**	**9.52**	**53.99**	**78.09**	22.71	48.89	**59.81**

数据设置：本文实验所用的多模态情感分析数据集 MELD（Multimodal Emotion Lines Dataset）[18]。MELD 数据集由 1432 个视频组成，可被切分成 13708 个片段，其中训练集 9989 个，验证集 1109 个和测试集 2610 个。MELD 中包含 7 种不同的情绪，分别是愤怒（*Anger*）、厌恶（*Disgust*）、恐惧（*Fear*）、喜悦（*Joy*）、中性（*Neutral*）、悲伤（*Sadness*）和惊讶（*Surprise*），每种情绪在数据集中的数量分布如表 1 所列。另外，MELD 中包含 3 种不同的情感，分别是正向（*Positive*）、负向（*Negative*）和中性（*Neutral*），

每种情感在数据集中的数量分布如表 2 所列。

参数设置和评价标准：关于实验中部分超参数的设置，我们设置 LSTM 的隐层单元数量为 256，批大小为 128，一共进行 30 次迭代。在实验中，我们根据 Poria 等[18]的研究工作选取加权平均的 F1 值（Weighted Average F1，记为 *w*-avg）作为综合评价标准，值越高代表性能越好。

3.2 基线方法

在实验中，我们将本文所提方法与一些最新的处理多模态情感分析的基线方法进行对比，具体在文本、语音以及多模态融合 3 个方面的实验性能进行比较。以下简要描述这些基线方法：

TFN：本方法由 Zadeh 等[5]提出，采用张量融合网络的方式，通过创建一个多维张量来捕捉多模态内容的单模态和多模态信息。

MFN：本方法由 Zadeh 等[12]提出，采用记忆融合网络的方式，通过使用多视图门控记忆单元来存储随时间变化的模态内部信息以及模态之间的交互信息。

BC-LSTM：本方法由 Poria 等[8]提出，采用上下文相关的多模态内容融合方式。该方法作为 MELD 数据集上提出的基线方法有着相当好的性能。

CMN：本方法由 Hazarika 等[19]提出，采用两层门控循环单元（GRU）来分别存储说话人信息以及当前多模态内容的上下文信息，通过融合双层 GRU 的输出来分析当前多模态内容所包含的情绪信息。

ICON：本方法由 Hazarika 等[13]提出，在 CMN 的基础上进行改进，增加一个全局门控循环单元（GlobalGRU）存储当前多模态内容的全局信息，进而对多模态内容进行情绪分析。

3.3 实验分析

不同方法在情感分类和情绪分类任务的实验结果如表 3 和表 4 所列。

在文本模态实验中，无论是在主任务还是辅助任务上，捕捉上下文信息的方法 **BC-LSTM**、**CMN** 和 **ICON** 表现均优于 **TFN** 和 **MFN**，以上的方法仅采用了主任务或辅助任务的文本特征。我们的方法在综合评价指标 *w*-avg 上表现最好，在情绪分类任务上比表现最佳的基线方法 **BC-LSTM** 高 0.94%，在情感分类任务上比表现最佳的 **CMN** 高 1.00%，表明文本模态的共享层的辅助表示可以同时提高主任务和辅助任务的性能。

在语音模态实验中，无论是在主任务还是辅助任务上，依然是捕捉上下文信息的方法比其他基线方法表现得更好，以上的方法仅采用了主任务或辅助任务的语音特征。由于语音模态特征的表征能力较弱，而且“*厌恶*”和“*恐惧*”类别的样本数目非常少，导致我们的模型无法预测出对应的类别，在个别类别上的表现无法达到最佳。不过，我们的方法在综合评价指标 *w*-avg 上表现得比所有基线方法好，在情绪任务上比表现最佳的方法 **ICON** 高 1.42%，在情感分类任务上比表现最佳的 **ICON** 高 1.29%，表明语音模态的共享层的辅助表示可以同时提高主任务和辅助任务的性能。

在文本和语音融合的多模态实验中，我们的方法在情绪分类任务上，虽然我们的方法未能在“*悲伤*”和“*惊讶*”上达到最佳，但是在其他五个类别以及 *w*-avg 上表现最佳，比所有基线方法分别高 3.99%、4.34%、3.83%、2.92%和 2.80%（*w*-avg）。在情感分类任务上所有类别均优于基线方法，在 *w*-avg 比五个基线方法分别高 4.02%、3.56%、2.04%、2.75%和 1.49%。实验结果表明，我们的方法采用主辅任务共享层的文本和语音辅助表示可以同时提升情绪和情感分类的性能。

4 结语

本文对基于多模态内容的情绪分类任务进行探索，首次引入一种多模态的联合学习框架，将多模态情绪分类作为主任务，多模态情感分类作为辅助任务，通过情感信息来辅助提升情绪分类任务的性能。本文方法中，我们采用私有网络层学习主任务的模态内部情绪信息，通过共享层学习主任务的辅助情绪表示以及辅助任务的完整情感表示，使用自注意力机制融合文本和语音模态信息以学习模态之间的交互信息。实验结果表明，我们的方法可以在多模态情感分析数据集上通过情感辅助信息大幅度提升情绪分类任务的性能，同时情感分类任务的性能也得到了一定的提升。

参考文献

[1] Morency L P, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th International Conference on Multimodal Interfaces. ACM, 2011: 169-176.

[2] Zadeh A, Zellers R, Pincus E, et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 2016, 31(6): 82-88.

[3] Gao Wei, Li Shoushan, Lee S Y M, et al. Joint learning on sentiment and emotion classification. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 1505-1508.

[4] Wang Rong, Li Shoushan, Zhou Guodong, et al. Joint sentiment and emotion classification with integer linear programming. International Conference on Database Systems for Advanced Applications. Springer, Cham, 2015: 259-265.

[5] Zadeh A, Chen Minghai, Poria S, et al. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250, 2017.

[6] Ma Dehong, Li Sujian, Wang Houfeng. Joint learning for targeted sentiment analysis. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4737-4742.

[7] Chen Ying, Hou Wenjun, Cheng Xiyao, et al. Joint learning for emotion classification and emotion cause detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 646-651.

[8] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 873-883.

[9] Pérez-Rosas V, Mihalcea R, Morency L P. Utterance-level multimodal sentiment analysis. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, 1: 973-982.

[10] Liu Zhun, Shen Ying, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064, 2018.

[11] Bertero D, Siddique F B, Wu C S, et al. Real-time speech emotion and sentiment recognition for interactive dialogue systems. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 1042-1047.

[12] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

[13] Hazarika D, Poria S, Mihalcea R, et al. ICON: Interactive conversational memory network for multimodal emotion detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2594-2604.

[14] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[15] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia. ACM, 2010: 1459-1462.

[16] Gu Yue, Yang Kangning, Fu Shiyu, et al. Hybrid attention based multimodal network for spoken language classification. Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2379-2390.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Poria S, Hazarika D, Majumder N, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508, 2018.

[19] Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in dyadic dialogue videos. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 2122-2132.