基于多任务学习的多模态情绪识别方法

吴良庆[1] 张 栋[1] 李寿山[1] 陈 珧[2]
(苏州大学计算机科学与技术学院 江苏 苏州 215006)[1]
(中国农业大学信息与电气工程学院 北京 100083)[2]

摘 要 情绪分析是自然语言处理的一项基本任务,目前在单模态信息(文本)上的研究已经相当成熟。但是对于包含文本、图像和语音3种模态信息的多模态内容(如视频)来说,额外增加的模态信息让情绪分析变得更具挑战性。为了提升多模态情绪识别任务的性能,文中提出了一种基于多任务学习的神经网络方法,该方法在考虑模态内部信息的同时,充分结合了3种模态之间的联系。具体而言,首先对3种模态信息进行预处理,得到相应的特征表示;其次,分别为每个模态构建私有的双向LSTM,从而获得单模态的内部信息;分别为两两组合(文本-图像、文本-语音和图像-语音)的双模态信息构建共享的双向LSTM层,以学习双模态之间的动态交互信息;接着,为3种模态组合的信息构建一个共享的双向LSTM,从而捕捉3种模态之间的动态交互信息;最后,把网络层中得到的单模态的内部信息和多模态的动态交互信息进行融合,通过全连接层和Sigmoid层获取最终的情绪识别结果。在单模态实验中,相比于目前的最佳方法,所提方法在文本、图像和语音3个方面对所有情绪识别的效果分别平均提高了6.25%,0.75%和2.38%;在多模态实验中,该方法在情绪识别任务中达到了平均65.67%的准确率,相比其他基准方法有了明显的提升。

关键词 多模态,情绪识别,多任务学习,自然语言处理
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.180901665

Multi-modal Emotion Recognition Approach Based on Multi-task Learning

WU Liang-qing[1] ZHANG Dong[1] LI Shou-shan[1] CHEN Ying[2]
(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)[1]
(College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)[2]

Abstract Emotion analysis is a fundamental task of natural language processing(NLP), and the research on single modality (text modality) has been rather mature. However, for multi-modal contents such as videos which consist of three modalities including text, visual and acoustic modalities, additional modal information makes emotion analysis more challenging. In order to improve the performance of emotion recognition on multi-modal emotion datasets, this paper proposed a neural network approach based on multi-task learning. This approach simultaneously considers both intra-modality and inter-modality dynamics among three modalities. Specifically, three kinds of modality information are first preprocessed to extract the corresponding features. Secondly, private bidirectional LSTMs are constructed for each modality to acquire the intra-modality dynamics. Then, shared bidirectional LSTMs are built for modeling inter-modality dynamics, including bi-modal (text-visual, text-acoustic and visual-acoustic) and tri-modal interactions. Finally, the intra-modality dynamics and inter-modality dynamics obtained in the network are fused to get the final emotion recognition results through fully-connected layers and the Sigmoid layer. In the experiment of uni-modal emotion recognition, the proposed approach outperforms the state-of-the-art by 6.25%, 0.75% and 2.38% in terms of text, visual and acoustic on average respectively. In addition, this approach can achieve average 65.67% in accuracy in multi-modal emotion recognition tasks, showing significant improvement compared with other baselines.

Keywords Multi-modal, Emotion recognition, Multi-task learning, Natural language processing

到稿日期:2018-09-06 返修日期:2018-12-01 本文受国家自然科学基金(61331011,61375073)资助。
吴良庆(1995-),男,硕士生,CCF学生会员,主要研究方向为自然语言处理、情感分析,E-mail:lqwu@stu.suda.edu.cn;张 栋(1991-),男,博士生,主要研究方向为自然语言处理、情感分析;李寿山(1980-),男,教授,主要研究方向为自然语言处理、情感分析;陈 珧(1978-),女,副教授,主要研究方向为自然语言处理、情感分析、信息抽取,E-mail:chenying@cau.edu.cn(通信作者)。

1　引言

随着微博、秒拍等新兴社交媒体的流行，越来越多的人喜欢在这些平台上通过分享视频表达自己对一些事物或者热点事件的观点和评论。这些包含用户情绪的多模态内容对于研究用户反馈、舆情发现和商品推荐等实际应用具有十分重要的作用。因此，面向多模态内容的情绪识别越来越受到学术界和工业界的关注。

多模态情绪识别[1-3]的主要任务是把基于文本的情绪识别[4]扩展为基于多模态信息的情绪识别，多模态的信息主要包含文本、图像和语音 3 种模态。多模态情绪识别的核心挑战是如何更好地对模态内部的信息(Intra-modality)以及模态之间的交互作用(Inter-modality)进行建模。模态内部的信息挖掘，主要是将各模态信息独立于其他的模态信息来单独进行处理，以提取模态内部的有用信息。与之对应的是模态之间具有交互作用的信息，情绪的表达通常是通过语言、视觉和声学行为之间的相互作用共同完成的，因此准确捕捉三者之间的联系至关重要。

目前，同时考虑模态内部信息以及模态之间的交互作用的研究方法(如 Zadeh 等[5]提出的张量融合方法，以及 Zadeh 等[6]提出的图记忆融合方法)都取得了不错的效果，但是这些方法的时间复杂度和空间复杂度都较高，尤其后者还需依赖于动态记忆网络方法。因而，本文将探索一种简单有效的方法来处理多模态情绪识别问题，把多个模态的信息看作多任务进行学习。众所周知，多任务学习方法在文本情感分析等领域表现得非常出色。Liu 等[7]采用了一种对抗性的多任务学习框架来减轻共享层和私有层潜在特征空间之间的相互干扰，从而提升文本分类的性能。Yu 等[8]提出一种一个基于卷积神经网络的方法，使用两个辅助任务来学习句子嵌入表示，以提高跨领域情感分析的性能。

受以上工作的启发，本文提出一种多任务融合学习网络(Multitask Fusion Learning Network，MFLN)方法来识别多模态情绪。具体而言，将文本信息、图像信息以及声音信息看成是多任务的输入：1)每种模态信息分别通过一个私有的双向 LSTM 层进行编码，以学习单个模态内部的变化信息；2)3 种模态信息之间两两结合形成 3 种组合，通过共享的双向 LSTM 层，以学习双模态之间的动态交互作用信息；3)联合 3 种模态的信息，经过同一个共享双向 LSTM 层来学习 3 种模态之间的动态联系。最后，把整个网络中学习到的多个模态的内部信息和模态之间的交互信息进行融合，以获取最终的情绪信息。实验结果表明，本文提出的方法可以显著提高多模态情绪识别的性能。

本文第 2 节介绍多模态情绪识别的相关工作；第 3 节详细描述本文提出的多任务融合学习方法；第 4 节介绍实验设置以及实验的结果和分析；最后给出结论并展望下一步的工作。

2　相关工作

多模态情绪识别作为多模态情感分析的一项基本任务，结合了语言以及非语言的信息去分析人们所表达的情感，现已成为了一个热点研究课题。最早的多模态情绪识别研究工作主要是在图像和语言模态信息上进行的。Glodek 等[9]联合使用面部表情以及语音信息进行多模态情绪识别研究；Ghosh 等[10]提出的方法也表明了联合视觉信息和听觉信息对情绪识别是有效的。

Morency 等[1]将多模态情感分析扩展到文本、图像和语音 3 种信息上，并且公开了第一个结合文本、图像和语音的多模态情感分析数据集 YouTube，自此有关多模态的研究便得到了迅速的发展。Wang 等[11]提出了 SAL-CNN 的方法来防止模型学习到依赖于视频中演讲者的特征，提升了模型在多模态情感分析中的泛化能力。Morency 等[1]通过 SVM 对产品和电影评论等多模态信息进行情感分析。但以上这些都是属于早融合(Early Fusion)的方法，会在信息输入时直接将多种模态信息进行拼接操作，这就可能导致模态内在的信息无法被有效利用，因此这种方法也被称为“特征级别的融合”(Feature-level Fusion)。与之对应的是晚融合(Late Fusion)的方法，该方法则是对每种模态信息单独做训练，考虑模态的内在信息，然后执行决策投票。Nojavanasghari 等[12]提出了先为每种模态信息训练一个模型，然后把模型结合起来做决策投票的方法。但是因为模态之间的交互作用比决策投票要复杂得多，所以这种方法难以学习到模态之间的相互联系。总而言之，在包含 3 种模态信息的数据集出现的早期，大部分研究多模态情感分析的工作没有同时考虑模态内部和模态之间的信息。

最近的研究工作主要考虑如何更好地把各模态的内在信息以及模态之间的交互作用结合起来。Zadeh 等[13]提出了记忆融合网络，通过多视图的门控记忆机制来存储模态内部以及模态之间的交互信息，从而实现多模态信息序列的同步。Chen 等[14]提出了一种二元门控机制，以消除模态信息中的噪声。实验结果表明了以上方法都取得了不错的效果；但正如前面所提到的，这些方法的实现较为复杂，而且依赖于动态的记忆网络。因此，本文提出了采用多任务融合学习的方法来研究多模态情绪识别的问题，通过私有与共享网络层的方式去考虑各模态内部信息以及模态之间的交互作用。

3　多任务融合学习方法

不同于相关工作中提到的方法，本文从多任务学习的角度探究了多模态情绪识别问题。图 1 为本文提出的多任务融合学习网络的模型框架图，该模型主要包含以下结构。

(1)特征输入层：将多模态内容中的文本、图像和语音特征输入到神经网络。

(2)Intra-modality 层：注重于单个模态(Uni-modal)内部的信息，采用的是私有的双向 LSTM 层。

(3)Inter-modality 层：包含两个部分，分别是双模态(Bi-modal)和三模态(Tri-modal)使用的共享双向 LSTM 层。

(4)预测分类层：分别使用单模态信息和多模态融合信息对数据集中的所有情绪类别进行识别，并把对每个情绪类别的识别都作为一个二分类任务，即对多个二分类任务进行预测。

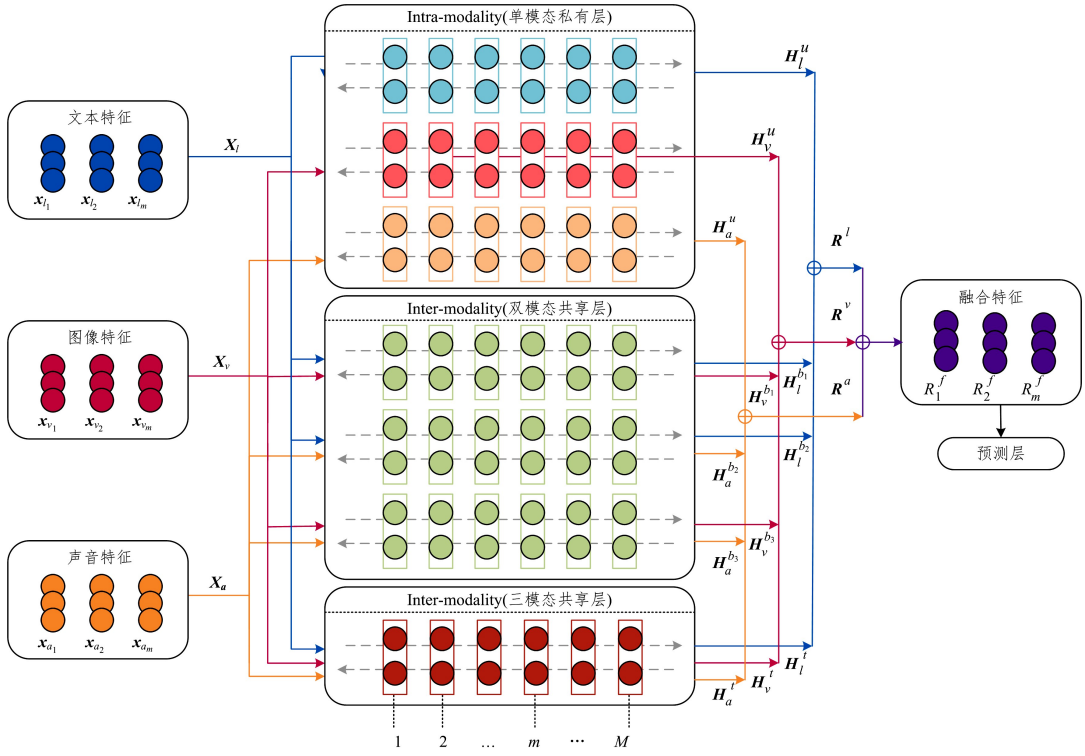


图 1 多任务融合学习网络的模型框架

Fig. 1 Overall architecture of multi-task fusion learning network

3.1 特征输入层

多个模态的时间序列必须对齐后才可以获得最佳性能[15]，如果每次的输入都需要做特征处理，那么这个工作量和时间的消耗都是巨大的。为了得到高效且可靠的数据加载，我们采用了卡内基梅隆大学提供的多模态数据 SDK[1] 去获取文本、图像以及语音的特征。文本的特征为 GloVe 词向量[16]。图像特征是通过把视频按每秒 30 帧的频率切割成图片，再由 FACET 面部表情分析框架得到[17]。语音特征则是通过把音频按每秒 100 帧的频率切分，再由 COVAREP 语音分析框架进行抽取[18]。3 种模态信息都含有对应的时间戳，在 SDK 中使用 P2FA 方法将文本与音频以及视频对齐[19]，即多帧的图片信息和语音信息最后会对应到相应的词的时间间隔。

3 种模态信息的特征如下所示：

$$N=\{l,v,a\} \tag{1}$$

$$\boldsymbol{X}_n=[\boldsymbol{x}_{n_m}:m\leqslant M,\boldsymbol{x}_{n_m}\in\mathbb{R}^{d_{x_n}},n=N] \tag{2}$$

其中，N 是 l(文本)、v(图像)和 a(语音)信息的集合，n 等于 N，即 n 包含 3 种模态信息。公式中的变量说明如表 1 所列。

表 1 公式中变量的说明

Table 1 Explanation of variables in formulas

变量名称	变量描述
N	文本、图像和语音模态的集合
n	N 的子集合
M	一句话的最大长度
$\boldsymbol{x}_{n_m}$	模态 n 的第 m 个词的特征向量
d_{x_n}	模态 n 的特征维度大小
$\boldsymbol{X}_n$	模态 n 中一句话的矩阵特征表示
d_{c_n}	LSTM 网络中隐层单元的数目

3.2 Intra-modality 层

Intra-modality 层主要关注的是单模态内部的信息，本文采用私有的双向 LSTM[20] 去获取单模态中上下文的语义信息。

对于单模态的内部信息，通过双向 LSTM 进行编码，具体表示如下：

$$\boldsymbol{h}^u_{n_m}=\overrightarrow{LSTM}(\boldsymbol{x}_{n_m})\oplus\overleftarrow{LSTM}(\boldsymbol{x}_{n_m}) \tag{3}$$

$$\boldsymbol{H}^u_n=[\boldsymbol{h}^u_{n_m}:m\leqslant M,\boldsymbol{h}^u_{n_m}\in\mathbb{R}^{2\times d_{c_n}}] \tag{4}$$

其中，$n=N$，表示分别考虑 3 种模态的内部信息；u 表示 Unimodal，即单模态的标记；$\boldsymbol{h}^u_{n_m}$ 是词 $\boldsymbol{x}_{n_m}$ 经过双向 LSTM 后的隐层表示；$\oplus$ 表示向量拼接的操作；$\boldsymbol{H}^u_n$ 是模态 n 中的一句话的矩阵表示。于是，可以分别得到单模态的文本表示 $\boldsymbol{H}^u_l$、图像表示 $\boldsymbol{H}^u_v$ 以及语音表示 $\boldsymbol{H}^u_a$。

3.3 Inter-modality 层

Inter-modality 层关注的是模态之间的交互作用，分为双模态和三模态之间的交互作用。本文采用共享的双向 LSTM 层去获取模态之间的交互信息。

3.3.1 双模态

在双模态中，由于文本、图像和语音 3 种模态信息可以两两组合，因此会产生 3 个共享的双向 LSTM 层，分别是文本与图像、文本与语音以及图像与语音的组合。

文本与图像的双模态组合的具体表示如下：

$$\boldsymbol{h}^{b_1}_{n_m}=\overrightarrow{LSTM}(\boldsymbol{x}_{n_m})\oplus\overleftarrow{LSTM}(\boldsymbol{x}_{n_m}) \tag{5}$$

$$\boldsymbol{H}^{b_1}_n=[\boldsymbol{h}^{b_1}_{n_m}:m\leqslant M,\boldsymbol{h}^{b_1}_{n_m}\in\mathbb{R}^{2\times d_{c_n}}] \tag{6}$$

其中，$n=N-\{a\}$，表示此共享的双向 LSTM 层只考虑了文本和图像两种模态信息；b_1 为文本和图像的双模态的标记；

1) https://github.com/A2Zadeh/CMU-MultimodalSDK

$\boldsymbol{h}_{n_m}^{b_1}$ 为词 $\boldsymbol{x}_{n_m}$ 经过双向 LSTM 后的隐层表示；$\boldsymbol{H}_n^{b_1}$ 为模态 n 中一句话的矩阵表示。于是，可以得到经过共享的双向 LSTM 层的文本表示 $\boldsymbol{H}_l^{b_1}$ 和图像表示 $\boldsymbol{H}_v^{b_1}$。

同样地，文本与语音的双模态组合的具体表示如下：

$$\boldsymbol{h}_{n_m}^{b_2}=\overrightarrow{LSTM}(\boldsymbol{x}_{n_m})\oplus\overleftarrow{LSTM}(\boldsymbol{x}_{n_m}) \tag{7}$$

$$\boldsymbol{H}_n^{b_2}=[\boldsymbol{h}_{n_m}^{b_2};m\leqslant M,\boldsymbol{h}_{n_m}^{b_2}\in\mathbb{R}^{2\times d_{c_n}}] \tag{8}$$

其中，$n=N-\{v\}$，表示此共享的双向 LSTM 层只考虑了文本和语音两种模态信息；b_2 为文本和语音双模态的标记；$\boldsymbol{h}_{n_m}^{b_2}$ 为词 $\boldsymbol{x}_{n_m}$ 经过双向 LSTM 后的隐层表示；$\boldsymbol{H}_n^{b_2}$ 为模态 n 中一句话的矩阵表示。于是，可以得到经过共享的双向 LSTM 层的文本表示 $\boldsymbol{H}_l^{b_2}$ 和语音表示 $\boldsymbol{H}_a^{b_2}$。

$$\boldsymbol{h}_{n_m}^{b_3}=\overrightarrow{LSTM}(x_{n_m})\oplus\overleftarrow{LSTM}(\boldsymbol{x}_{n_m}) \tag{9}$$

$$\boldsymbol{H}_n^{b_3}=[\boldsymbol{h}_{n_m}^{b_3};m\leqslant M,\boldsymbol{h}_{n_m}^{b_3}\in\mathbb{R}^{2\times d_{c_n}}] \tag{10}$$

其中，$n=N-\{l\}$，表示此共享的双向 LSTM 层只考虑图像和语音两种模态信息；b_3 为文本和语音双模态的标记；$\boldsymbol{h}_{n_m}^{b_3}$ 为词 $\boldsymbol{x}_{n_m}$ 经过双向 LSTM 后的隐层表示；$\boldsymbol{H}_n^{b_3}$ 为模态 n 中一句话的矩阵表示。于是，可以得到经过共享的双向 LSTM 层的图像表示 $\boldsymbol{H}_v^{b_3}$和语音表示 $\boldsymbol{H}_a^{b_3}$。

3.3.2　三模态

在三模态的交互中，我们将文本、图像和语音 3 种模态信息同时输入到一个共享的双向 LSTM 网络层，具体表示如下：

$$\boldsymbol{h}_{n_m}^{t}=\overrightarrow{LSTM}(\boldsymbol{x}_{n_m})\oplus\overleftarrow{LSTM}(\boldsymbol{x}_{n_m}) \tag{11}$$

$$\boldsymbol{H}_n^{t}=[\boldsymbol{h}_{n_m}^{t};m\leqslant M,\boldsymbol{h}_{n_m}^{t}\in\mathbb{R}^{2\times d_{c_n}}] \tag{12}$$

其中，$n=N$，表示此共享层把 3 种模态的内部信息都进行了考虑；t 表示 Tri-modal，即三模态的标记；$\boldsymbol{h}_{n_m}^{t}$ 为词 $\boldsymbol{x}_{n^m}$ 经过双向 LSTM 后的隐层表示；$\boldsymbol{H}_n^{t}$ 为模态中一句话的矩阵表示。于是，可以分别得到经过共享层作用的 3 种模态的表示 $\boldsymbol{H}_l^{t}$，$\boldsymbol{H}_v^{t}$ 和 $\boldsymbol{H}_a^{t}$。

3.4　预测分类层

预测分类层，分别使用单模态信息和多模态融合信息对多个情绪识别任务进行预测。

为了得到文本、图像和语音信息的完整表示，我们首先将 Intra-modality 层和 Inter-modality 层学习到的所有文本、图像和语音表示分别融合在一起，具体表示如下：

$$\boldsymbol{R}^{l}=\boldsymbol{H}_l^{u}\oplus\boldsymbol{H}_l^{b_1}\oplus\boldsymbol{H}_l^{b_2}\oplus\boldsymbol{H}_l^{t} \tag{13}$$

$$\boldsymbol{R}^{v}=\boldsymbol{H}_v^{u}\oplus\boldsymbol{H}_v^{b_2}\oplus\boldsymbol{H}_v^{b_2}\oplus\boldsymbol{H}_v^{t} \tag{14}$$

$$\boldsymbol{R}^{a}=\boldsymbol{H}_a^{u}\oplus\boldsymbol{H}_a^{b_2}\oplus\boldsymbol{H}_a^{b_3}\oplus\boldsymbol{H}_a^{t} \tag{15}$$

其中，$\boldsymbol{R}^{l}$ 是把经过 Intra-modality 层作用的 $\boldsymbol{H}_l^{u}$、通过双模态共享层作用的 $\boldsymbol{H}_l^{b_1}$ 和 $\boldsymbol{H}_l^{b_2}$ 以及经过三模态共享层作用的 $\boldsymbol{H}_l^{t}$ 结合起来，得到的文本模态的完整表示；$\oplus$ 表示向量拼接的操作。同样地，可以得到图像模态的完整表示 $\boldsymbol{R}^{v}$ 和语音模态的完整表示 $\boldsymbol{R}^{a}$。

接下来，我们采取直接拼接 3 种模态信息的方法得到模态特征的融合表示：

$$\boldsymbol{R}^{f}=\boldsymbol{R}^{l}\oplus\boldsymbol{R}^{v}\oplus\boldsymbol{R}^{a} \tag{16}$$

其中，$\boldsymbol{R}^{f}$ 就是 3 种模态信息的融合表示，即把 Intra-modality 层和 Inter-modality 层得到的单模态、双模态以及三模态的表示融合在一起。在最终分类前，我们需要对 $\boldsymbol{R}^{f}$ 进行一个降维操作，采用最大值池化的方法，$\boldsymbol{R}^{F}=\max\text{ pooling}\{\boldsymbol{R}^{f}\}$，以保留最强的特征信息。

3.4.1　单模态分类

在分别得到文本、图像和语音的完整表示后，使用激活函数 tanh 和 sigmoid 层进行单模态信息的情绪识别预测，具体表示如下：

$$\hat{y}^{n}=\sigma(\boldsymbol{W}_q\cdot(\tanh(\boldsymbol{W}_p\cdot\boldsymbol{R}^{n}+\boldsymbol{b}_p))+\boldsymbol{b}_q) \tag{17}$$

其中，$n=N$；$\hat{y}^{l}$，$\hat{y}^{v}$ 和 $\hat{y}^{a}$ 分别是单独使用文本、图像和语音模态信息得到的情绪识别结果；$\boldsymbol{W}_p$ 和 $\boldsymbol{b}_p$ 是 tanh 层的权重和偏置；$\boldsymbol{W}_q$ 和 $\boldsymbol{b}_q$ 是 sigmoid 层的权重与偏置。

3.4.2　多模态分类

同样地，在得到了融合特征的表示 $\boldsymbol{R}^{F}$ 后，使用激活函数 tanh 和 sigmoid 层进行最终的预测，具体表示如下：

$$\hat{y}^{F}=\sigma(\boldsymbol{W}_q\cdot(\tanh(\boldsymbol{W}_p\cdot\boldsymbol{R}^{F}+\boldsymbol{b}_p))+\boldsymbol{b}_q) \tag{18}$$

其中，$\hat{y}^{F}$ 是文本、图像和语音 3 种模态信息融合后得到的情绪识别结果；$\boldsymbol{W}_p$，$\boldsymbol{b}_p$，$\boldsymbol{W}_q$ 和 $\boldsymbol{b}_q$ 的含义同上。

3.5　优化策略

在模型训练的过程中，无论是单模态分类还是多模态分类，均选择交叉熵误差作为损失函数，公式如下：

$$Loss(\hat{y},y)=-\sum_{s=1}^{S}\sum_{c=1}^{C}y_s^{c}\cdot\log\hat{y}_s^{c} \tag{19}$$

其中，y 是真实标签，$\hat{y}$是模型预测的概率，S 是训练样本的总数，C 是类别的数目。同时，实验中采用 Adadelta 优化器来优化模型的参数[21]。

4　实验

本节将系统地分析所提方法在多模态情绪识别上的效果。

4.1　实验设置

本文实验所用的多模态情绪识别数据集是由卡内基梅隆大学提供的 MOSEI(Multimodal Opinion Sentiment and Emotion Intensity)数据集，本文的工作主要是对其中的情绪识别任务进行研究。

MOSEI 数据由 3 228 个视频组成，总计 23 453 个句子。由于该数据集本身的原因，如果样本中的 3 种模态信息存在丢失情况，则舍弃该样本，最终保留了 22 634 个样本用例。MOSEI 数据集中包含了 6 种不同的情绪，分别是愤怒(Anger)、厌恶(Disgust)、恐惧(Fear)、高兴(Happy)、悲伤(Sad)和惊讶(Surprise)，我们把情绪强度值为 0 的样本视为不包含该情绪的负例，情绪强度值大于 0 的即为句子中存在该情绪的正例。统计得到 6 种情绪类别的正例占比分布，如表 2 所列，可以看出：情绪类别的样本分布都是不平衡的。因此我们对每一种情绪类别的训练集和验证集都做了降采样[22]的处理，使得训练集和验证集内的正负例各占一半，测试集不变，得到的每一个类别的样本数量如表 3 所列。由于一个句子中可能包含多种情绪，因此还需要考虑多标签的问题[23]。本文把多标签任务拆分成对每一种情绪类别单独做二元分类的任务，相当于将多标签任务转换成了 6 个情绪分类的任务，这也是在 3.4 节中采用激活函数 sigmoid 做最后的二分类预测的原因。

表 2 各情绪类别正例样本的占比分布

Table 2 Distribution of positive samples in each emotion category

类别	愤怒	厌恶	恐惧	高兴	悲伤	惊讶
训练集	0.22	0.18	0.08	0.54	0.26	0.10
验证集	0.18	0.15	0.08	0.53	0.28	0.11
测试集	0.23	0.17	0.08	0.54	0.24	0.09

表 3 经过降采样后各情绪类别的训练集、验证集和测试集数目

Table 3 Number of training, valid and test samples for each emotion category after under-sampling

类别	训练集	验证集	测试集	总样本数
愤怒	6998	666	4614	12278
厌恶	5882	558	4614	11054
恐惧	2612	326	4614	7552
高兴	15052	1712	4614	21378
悲伤	8454	1020	4614	14088
惊讶	3258	388	4614	8260

由于情绪类别的测试样本内正负样例不平衡，我们采用了 Tong 等[24]提出的加权的准确率(Weighed Accuracy, WA)，其计算公式如下：

$$WA=\frac{TP\times N/P+TN}{2N} \tag{20}$$

其中，TP 是预测为正且实际为正的样例；相应地，TN 是预测为负且实际也为负的样例；而 P 和 N 分别表示正向样例和负向样例的数量。相比于不加权的准确率，在样本不平衡的情况下，加权的准确率更能提供信息量[6]。

4.2 实验结果

我们将所提方法在文本、图像、语音以及三模态融合 4 个方面与目前最佳(state of the art)的方法进行了对比，SOTA1 和 SOTA2 分别表示在各模态上当前效果最佳和第二的方法，实验结果如表 4 所列。

表 4 所提方法与目前最佳方法的比较

Table 4 Comparison between the proposed approach and other SOTA approaches

类别		愤怒	厌恶	恐惧	高兴	悲伤	惊讶
文本	SOTA2	56.0$^{\cup}$	59.0$^{\square}$	56.2$^{\square}$	53.0*	53.8†	53.2$^{\times}$
	SOTA1	56.6†	64.0*	58.8$^{\times}$	54.0$^{\square}$	54.0$^{\square}$	54.3*
	MFLN	**63.1**	**71.0**	**60.6**	**64.6**	**60.0**	**59.9**
图像	SOTA2	54.4†	54.4$^{\heartsuit}$	51.3	53.4†	54.3*	51.3*
	SOTA1	**60.0**$^{\square}$	**60.3**†	**64.2**$^{\heartsuit}$	57.4$^{\bullet}$	57.7$^{\square}$	51.8$^{\square}$
	MFLN	58.7	59.7	56.9	**69.2**	**59.2**	**52.2**
语音	SOTA2	55.5$^{\bigstar}$	58.9*	58.5*	57.2$^{\cap}$	58.9$^{\bigstar}$	52.2$^{\heartsuit}$
	SOTA1	56.4$^{\triangle}$	60.9$^{\square}$	**62.7**$^{\square}$	61.5$^{\square}$	**62.0**$^{\cap}$	**54.3**$^{\bigstar}$
	MFLN	**63.4**	**69.8**	62.4	**61.8**	60.9	53.8
多模态	SOTA2	60.5$^{\diamond}$	67.0$^{\flat}$	60.0$^{\heartsuit}$	66.3$^{\clubsuit}$	59.2$^{\square}$	53.3$^{\sharp}$
	SOTA1	62.6$^{\clubsuit}$	69.1$^{\clubsuit}$	62.0$^{\clubsuit}$	66.5$^{\diamond}$	60.4$^{\clubsuit}$	53.7$^{\clubsuit}$
	MFLN	**66.0**	**72.3**	**62.5**	**69.2**	**63.0**	**61.0**

以下将简单描述这些方法，括号中的符号与表中结果对应。

SVM(♡)：一种采用早融合方法对多模态特征训练的 SVM 模型[25]。

DAN(♡)：一种采用词的分布表示的深度平均网络[26]，主要用于情感分析任务。

RF(●)：一种采用随机森林[27]作为多模态情绪识别任务中非神经网络分类器的方法。

Adieu-Net(△)：一种采用端到端的语音情绪识别方法[28]，通过结合 CNN 与 LSTM 网络来自动学习语音信号的最佳表示。

DF(♭)：一种采用深度融合的方法[29]，为每种模态训练一个深度模型，对每个模态网络的输出进行决策投票。

EF-LSTM(□)：一种采用早期融合 LSTM[20]的方法，在每个时间步连接来自不同模态的输入，并将其用作单个 LSTM 的输入。

SER-LSTM(∩)：一种在音频谱图的卷积运算上使用递归神经网络的模型[30]。

CNN-LSTM(∪)：一种在每个时间戳执行面部区域卷积的循环模型[31]，并把得到的输出作为 LSTM 的输入。

DynamicCNN(×)：一种用于句子语义建模的卷积结构，是基于文本的情感分析中最先进的模型之一[32]。

TFN(◇)：一种采用张量融合网络的方法，通过创建一个多维张量，在多模态数据中学习单模态、双模态和三模态的信息[4]。

MFN(♯)：一种采用记忆融合网络的方法，通过使用多视图门记忆来存储随时间变化的各模态内部信息以及模态之间的交互作用[13]。

GMFN(♣)：一种采用图形记忆进行融合的网络[6]，将动态融合图与记忆融合网络相结合。

DHN(†)：一种采用深度高速公路网络[33]的方法，能有效地减缓训练深层神经网络的梯度消失问题，可使深层神经网络不再仅仅具有浅层神经网络的效果。

RHN(★)：一种在循环神经网络内部引入高速公路层[34]的方法，保留了 LSTM 的易训练性。

MFLN：本文提出的多任务融合学习网络。

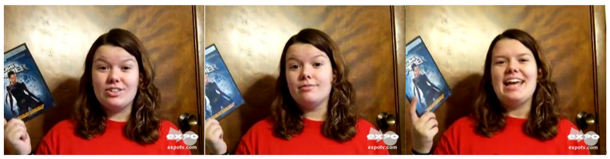
从表 4 中可以发现：在单模态的实验中，以文本模态为例，当文本信息经过 Inter-modality 层的共享层时，其他两种模态信息的训练会影响到文本的表示，从而可以增强文本信息的泛化表征能力。图像和语音模态信息亦如此。对比目前的最佳性能(SOTA1)，MFLN 在使用文本信息时，在每种情绪上的结果(WA)平均可以提高 6.25%，图像模态在高兴、悲伤和惊讶类别上平均可以提高 4.56%，语音模态在愤怒、厌恶和高兴类别上平均可以提高 5.4%。虽然在图像和语音模态上本文方法在有些情绪上的识别效果略逊于 SOTA1，但是该方法在各个类别上均比 SOTA2 表现得好。

在多模态实验中，由于 EF-LSTM 和 SVM 采用了早融合的方法，模态内在的信息无法被有效利用，使得这些方法的结果不尽理想；DF 只在厌恶类别上达到目前第二好的效果，因为 DF 采用的是晚融合方法，难以学习到模态之间的交互作用。GMFN，MFN 和 TFN 这类方法都考虑了各模态内部信息以及模态之间的交互作用，是目前处理多模态问题的主流方法。其中，GMFN 的表现是目前最佳的(除了在高兴类别上由 TFN 领先)。对比目前性能最佳的 SOTA1，所提方法 MFLN 在所有情绪类别的识别性能上可以平均提高 3.28%，这表明我们的多任务融合学习网络可以更有效地学习各模态

内部信息以及模态之间的交互作用，对有效的情绪信息进行融合，以识别情绪，这表明模态之间的动态交互信息在多模态任务中是相当重要的一环。

4.3　例子分析

本节通过图 2 所示的例子来进一步说明所提方法的优越性，预测结果如表 5 所列。在使用单模态信息预测时，通过文本信息，厌恶的情绪是比较容易判别的，但是无法识别出高兴的情绪。图像信息正好相反，通过图像可以识别出主角露出的笑容，而厌恶的表情在视频中始终表现得不明显，因此图像信息能识别出高兴的情绪，却无法正确预测出厌恶的情绪。对于语音信息，主角说话时是带有一点对光盘价格高的嫌弃语气的，因此其仅预测出了厌恶的情绪。在采用了多模态信息融合后，模型综合考虑了 3 种模态信息之间的动态交互作用，所提方法识别出了正确的情绪类别。



"If you can find it for cheap and you like Tomb Raider, that would be a good thing. But don't pay over 5 dollars for this."（如果能找到《古墓丽影》便宜的片源而且你也喜欢这部电影的话，那会是非常好的。但是不要以高于 5 美元的价格买这部电影的光盘）

语音：

A bit disgust at the high price of the DVD.（语气中带有厌恶光盘的高价格）

图 2　多模态情绪识别的例子

Fig. 2　Example of multi-modal emotion recognition

表 5　所提方法的预测结果与真实标签的比较

Table 5　Comparison between predictions of proposed approach and ground truth

类别	文本	图像	语音	多模态	真实标签
厌恶	1	0	1	1	1
高兴	0	1	0	1	1

结束语　本文提出了从多任务学习的角度去探究多模态情绪识别的方法，模型通过私有层学习到单模态内部的信息，再将其与通过共享层学习到的模态之间的交互信息结合起来，去捕捉对情绪识别任务有效的信息。实验结果表明：本文提出的方法能够提升多模态情绪识别的性能。

在接下来的工作中，我们将考虑把我们的方法运用到其他多模态的数据集上，以验证方法的有效性。另外，我们也会对模型进行进一步的探索，尝试加入注意力机制，更好地判别出各种模态信息中对情绪识别有效的部分。

参 考 文 献

[1] MORENCY L P, MIHALCEA R, DOSHI P. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web [C]// Proceedings of 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[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.

[3] PORIA S, CAMBRIA E, GELBUKH A F. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2015: 2539-2544.

[4] HUANG L, LI S S, ZHOU G D. Emotion recognition of Chinese microblogs with syntactic information [J]. Computer Science, 2017, 44(2): 244-249. (in Chinese)
黄磊，李寿山，周国栋. 基于句法信息的微博情绪识别方法研究[J]. 计算机科学，2017，44(2)：244-249.

[5] ZADEH A, CHEN M, PORIA S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2017: 1103-1114.

[6] ZADEH A, LIANG P P, VANBRIESEN J, et al. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph[C]// Proceedings of the Meeting of the Association for Computational Linguistics. 2018: 2236-2246

[7] LIU P, QIU X, HUANG X. Adversarial Multi-task Learning for Text Classification[C]// Proceedings of the Meeting of the Association for Computational Linguistics. 2017: 1-10.

[8] YU J, JIANG J. Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2016: 236-246.

[9] GLODEK M, TSCHECHNE S, LAYHER G, et al. Multiple Classifier Systems for the Classification of Audio-Visual Emotional States[C]// Proceedings of International Conference on Affective Computing and Intelligent Interaction. Springer-Verlag, 2011: 359-368.

[10] GHOSH S, LAKSANA E, MORENCY L P, et al. Representation Learning for Speech Emotion Recognition[C]// Proceedings of INTERSPEECH. 2016: 3603-3607.

[11] WANG H, MEGHAWAT A, MORENCY L P, et al. Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis[J]. arXiv: 1609.05244

[12] NOJAVANASGHARI B, HUGHES C E, MORENCY L P. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children[C]// Proceedings of International Conference on Multimodal Interaction. ACM, 2016: 137-144.

[13] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory Fusion Network for Multi-view Sequential Learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[14] CHEN M, WANG S, LIANG P P, et al. Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning [C]// Proceedings of International Conference on Multimodal Interaction. ACM, 2017: 163-171.

[15] ZADEH A,LIANG P P,PORIA S,et al. Multi-attention Recurrent Network for Human Communication Comprehension[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[16] PENNINGTON J,SOCHER R,MANNING C. Glove:Global Vectors for Word Representation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[17] EKMAN P. An argument for basic emotions. [J]. Cognition & Emotion,1992,6(3/4):169-200.

[18] DEGOTTEX G,KANE J,DRUGMAN T,et al. COVAREP — A Collaborative Voice Analysis Repository for Speech Technologies[C]// Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2014:960-964.

[19] YUAN J,LIBERMAN M. Speaker Identification on the SCOTUS Corpus[J]. Journal of the Acoustical Society of America,2008,123(123):3878.

[20] HOCHREITER S,SCHMIDHUBER J. Long Short-term Memory. [J]. Neural Computation,1997,9(8):1735-1780.

[21] ZEILER M D. ADADELTA:An Adaptive Learning Rate Method[J]. arXiv:1212.5701.

[22] YIN H,LI S S,GONG Z X,et al. Imbalanced Emotion Classification Based on Multi-channel LSTM[J]. Journal of Chinese Information Processing,2018,32(1):139-145. (in Chinese)
殷昊,李寿山,贡正仙,等. 基于多通道 LSTM 的不平衡情绪分类方法[J]. 中文信息学报,2018,32(1):139-145.

[23] HUANG Y,WANG W,WANG L,et al. Multi-task Deep Neural Network for Multi-label Learning[C]// Proceedings of IEEE International Conference on Image Processing. IEEE,2014:2897-2900.

[24] TONG E,ZADEH A,JONES C,et al. Combating Human Trafficking with Multimodal Deep Models[C]// Proceedings of the Meeting of the Association for Computational Linguistics. 2017:1547-1556.

[25] CORTES C,VAPNIK V. Support-vector Networks[J]. Machine Learning,1995,20(3):273-297.

[26] IYYER M,MANJUNATHA V,BOYD-GRABER J,et al. Deep Unordered Composition Rivals Syntactic Methods for Text Classification[C]// Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. 2015:1681-1691.

[27] HO T K. The Random Subspace Method for Constructing Decision Forests[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,1998,20(8):832-844.

[28] TRIGEORGIS G,RINGEVAL F,BRUECKNER R,et al. Adieu features? End-to-end Speech Emotion Recognition Using a Deep Convolutional Recurrent Network[C]// Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2016:5200-5204.

[29] NOJAVANASGHARI B,GOPINATH D,KOUSHIK J,et al. Deep Multimodal Fusion for Persuasiveness Prediction[C]// Proceedings of International Conference on Multimodal Interaction. ACM,2016:284-288.

[30] LIM W,JANG D,LEE T. Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks[C]// Proceedings of Signal and Information Processing Association Summit and Conference. IEEE,2017:1-4.

[31] KAHOU S E,MICHALSKI V,KONDA K,et al. Recurrent Neural Networks for Emotion Recognition in Video[C]// Proceedings of International Conference on Multimodal Interaction. ACM,2015:467-474.

[32] KALCHBRENNER N,GREFENSTETTE E,BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[J]. arXiv:1404.2188.

[33] SRIVASTAVA R K,GREFF K,SCHMIDHUBER J. Training Very Deep Networks[J]. arXiv:1507.06228.

[34] ZILLY J G,SRIVASTAVA R K,KOUTNÍk J,et al. Recurrent Highway Networks[J]. arXiv:1607.03474.