北京大学学报(自然科学版) 第 57 卷 第 1 期 2021 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 57, No. 1 (Jan. 2021)
doi: 10.13209/j.0479-8023.2020.085

一种基于多任务学习的多模态情感识别方法

林子杰[1] 龙云飞[2] 杜嘉晨[1] 徐睿峰[1,†]

1. 哈尔滨工业大学(深圳)计算机科学与技术学院, 深圳 518055; 2. School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ; † 通信作者, E-mail: xuruifeng@hit.edu.cn

摘要 为了通过设置辅助任务学习到更具有情感倾向性的视频和语音表示, 进而提升模态融合的效果, 提出一种基于多任务学习的多模态情感识别模型, 使用多模态共享层来学习视觉和语音模型的情感信息。在MOSI数据集和MOSEI数据集上的实验表明, 添加两个辅助的单模态情感识别任务后, 模型可以学习到更有效的单模态情感表示, 并且在两个数据集上的情感识别准确率比目前性能最佳的单任务模型分别提升0.8%和2.5%。

关键词 多模态信息; 情感识别; 模态融合; 多任务学习

A Multi-modal Sentiment Recognition Method Based on Multi-task Learning

LIN Zijie[1], LONG Yunfei[2], DU Jiachen[1], XU Ruifeng[1,†]

1. School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055;
2. School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ;
† Corresponding author, E-mail: xuruifeng@hit.edu.cn

Abstract In order to learn more emotionally inclined video and speech representations through auxiliary tasks, and improve the effect of multi-modal fusion, this paper proposes a multi-modal sentiment recognition method based on multi-task learning. A multimodal sharing layer is used to learn the sentiment information of the visual and acoustic modes. The experiment on MOSI and MOSEI data sets shows that adding two auxiliary single-modal sentiment recognition tasks can learn more effective single-modal sentiment representations, and improve the accuracy of sentiment recognition by 0.8% and 2.5% respectively.

Key words multi-modal information; sentiment recognition; multi-modal fusion; multi-task learning

在人类情感交流中, 每个人作为个体, 通过聆听语言、观察表情以及分析语言内容等方式, 感受其他人的情感变化, 识别情感状态信息, 进而进行情感交流。如果想让模型如同人类一样理解情感, 就需要对人类多种情感的表达(视觉、语音和文本)进行识别, 让机器具有捕捉多模态情感特征并进行处理, 最后表达出相应人类情感的能力。

目前, 大多数关于情感识别模型的研究集中在语言(尤其是文本)模态上, 但是单模态文本情感识别存在识别率不够高和鲁棒性差等缺点。多模态情感识别可以有效地利用多种模态识别包含的信息, 捕捉模态之间的互补信息, 从而提升模型的识别能力和泛化能力。在进行模态融合之前, 若能够更好地挖掘视觉和语音模态的情感倾向特征, 则3种模态表示之间的任务相关性更强, 也更有助于模态的融合。

在多模态情感分析领域, 已经提出大量计算模型, 包括张量融合网络[1]、记忆融合网络[2]和多级注意力循环网络[3]等。传统的多模态情感分析模型通常将单个模态信号建模为独立的向量表示, 通过模态融合, 进行模态之间相互关联的建模, 但是在模态融合前, 缺少对情感特征的提取, 导致模态间

国家自然科学基金(61876053, 61632011, 62006062)、深圳市基础研究学科布局项目(JCYJ20180507183527919, JCYJ20180507183608379)和广东省新冠肺炎疫情防控科研专项(2020KZDZX1224)和深圳市技术攻关项目(JSGG20170817140856618)资助
收稿日期: 2020–06–08; 修回日期: 2020–08–14

的共享情感特征不易被识别。为了解决这一问题，Akhtar 等[4]提出使用多任务学习框架，对情绪识别任务和情感识别任务间的关联建模，通过相关任务之间的关联性，对不同模态中的情感特征进行提取。但是，该方法未考虑不同模态信息情感表达程度的不同，可能导致模态融合效果不明显，且难以解释模态之间的关联性。

为解决传统的基于多任务学习的多模态情感识别模型中的问题，本文提出一种不需要额外情绪标注的，适用于多模态情感识别任务的多任务学习框架，通过引入单模态情感识别任务，可以学习到更具有情感倾向性的视频和语音表示，进而提升模态融合的效果。

1 相关工作

1.1 多模态情感识别

Baltrušaitis 等[5]将多模态机器学习的研究分为模态表示、模态传译、模态对齐、模态融合和合作学习 5 个方面，多模态情感识别研究主要涉及模态表示、模态对齐、模态融合和合作学习 4 个方面，当前多集中在模态融合层面。

模态融合的目的是将不同单模态中提取的信息整合到一个紧凑的多模态表示中[6]。根据融合发生的阶段，分为早期融合、晚期融合和混合融合。早期融合[7]指在编码前对多模态的特征进行融合，是特征层面的融合。由于发生在特征提取阶段，早期融合能够有效地提取模态间的交互信息，但可能忽略单模态内的交互信息。较典型的早期融合模型是 EF-LSTM[3]，该模型将文本、语音和图像 3 种模态的特征表示进行拼接，得到多模态表示，再输入 LSTM 中进行编码。晚期融合[7]发生在解码之后，是决策层面上的融合，能够提取模态内的交互信息，但无法提取模态间的交互信息，常用的方法有平均[8]、投票[9]和加权[10]等。混合融合则组合了前两种融合方法。由于深度学习方法主要用于特征层的处理，基于深度学习的模态融合方法大多采用早期融合策略和混合策略。本文主要针对早期融合方法进行研究。

1.2 基于多模态偏移门的模态融合方法

Rahman 等[11]提出的 M-BERT 模型将预训练模型应用在多模态情感识别任务中。与 BERT 不同，M-BERT 在输入层与编码层之间加入模态融合层，并使用多模态偏移门限机制[12](multimodal shifting gate，MSG)，实现 3 种模态的融合。MSG 通过将词向量分别与视觉、语音模态的特征向量拼接，用于产生两个模态的门向量，作为模态融合的权重，生成偏移向量。偏移向量乘上一个比例因子后与词向量相加，得到修正后的多模态词向量。

1.3 多任务学习

多任务学习(multi-task learning, MTL)是机器学习的一个子领域，其训练过程中包含多个学习任务，通过利用不同任务间的共性和差异来提高模型的泛化能力和预测准确率[13−15]。一般来说，训练不同种类任务需要不同的模型结构，要实现多任务学习，就需要实现模型间的参数共享。因此，多任务学习模型是由多个结构重叠的机器学习模型的组合，重叠的部分是多个学习任务在反向传播过程中都必须经过的，称为共享层(shared layers)。

多任务学习模型的参数共享策略主要有硬共享[16](hard sharing)和软共享[17](soft sharing)两种，其次还有分层共享(hierarchical sharing)和稀疏共享[18](sparse sharing)等。硬共享是最常见的共享策略，不同任务共享除输出层外的模型部分。硬共享可以同时训练多个任务的通用表示，有效地避免由于训练数据较少导致的过拟合风险。软共享策略不直接共享模型结构，每个任务都有自己的模型和参数，通过对模型相似部分的参数进行正则化[17,19]来保证模型的参数相似性。

2 基于多任务学习的多模态情感识别方法

本文基于多任务学习的多模态情感识别模型框架如图 1 所示，模型由以下 3 个部分组成。

1) 多模态任务共享层：包括 3 个任务模型共享的部分，用于学习视频和语音表示，位于输入层之后，编码层之前。在训练的过程中，每一次反向传播都会经过共享层。

2) 多模态情感识别模型：是加入了共享层的 M-BERT，除共享层外的部分，只有在其输入为 3 种模态的特征向量时，才会在反向传播过程中更新参数。

3) 单模态情感识别模型：即视频/语音情感识别任务模型，包括输入层、共享层、编码层和预测层。除共享层外，只有在输入是任务对应模态的特征向量时，才会在反向传播过程中更新参数。

2.1 多模态任务共享层

我们在输入层后面加入视觉和语音共享层，用于学习更适合情感分类任务的视觉/语音表示。图1中的视觉隐向量和声学隐向量为视觉特征向量 $\boldsymbol{V}_i$ 和声学特征向量 $\boldsymbol{A}_i$ 经过共享层后的输出。这里为视频和语音模态分别设置一个线性层作为共享层，共享层输出的视觉/声学隐向量 $\boldsymbol{H}_i^v$ 和 $\boldsymbol{H}_i^a$：

$$\boldsymbol{H}_i^v = \boldsymbol{W}_v \cdot \boldsymbol{V}_i + \boldsymbol{b}_v, \tag{1}$$

$$\boldsymbol{H}_i^a = \boldsymbol{W}_a \cdot \boldsymbol{A}_\mathrm{i} + \boldsymbol{b}_a, \tag{2}$$

其中，$i=1,2,\ldots,N$，$\boldsymbol{H}_i^v \in \mathbb{R}^{\mathrm{d_v}}$，$\boldsymbol{H}_i^a \in \mathbb{R}^{d_a}$，$\boldsymbol{W}_v$，$\boldsymbol{W}_a$，$\boldsymbol{b}_v$ 和 $\boldsymbol{b}_a$ 分别为视频和语音模态共享层的参数权重和偏置，$\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d_v}$，$\boldsymbol{W}_a \in \mathbb{R}^{d_a \times d_a}$，$\boldsymbol{b}_v \in \mathbb{R}^{d_v}$，$\boldsymbol{b}_a \in \mathbb{R}^{d_a}$。当模型的输入为多模态数据时，进行多模态情感识别训练，将共享层输出的视觉隐向量和声学隐向量传入MSG单元，与词向量一起进行模态融合；当输入仅为视频/语音模态的数据时，进行单模态情感识别训练，学习到的视觉/声学隐向量将传入后续的单模态编码器中，经过预测层输出情感极性。

2.2 多模态情感识别模型

本文使用加入多模态任务共享层的M-BERT模型作为多模态情感识别模型，共享层的位置在特征输入层与模态融合层之间。模型将长度为 N 的词序列(L_1, L_2, …, L_N)、视觉特征序列(V_1, V_2, …, V_N)和声学特征序列(A_1, A_2, …, A_N)作为输入，词序列经BERT输入层映射为词嵌入序列(E_1, E_2, …, E_N)。多模态情感识别模型的输出为预测的情感得分 $\tilde{y}$，计算真实情感得分 y 和预测情感得分 $\tilde{y}$ 之间的平均绝对误差 $\mathcal{L}_\mathrm{m}$：

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N} | y_i - \tilde{y}_i | 。 \tag{3}$$

2.3 单模态情感识别模型

单模态情感识别模型如图2所示，使用双向LSTM网络作为单模态编码器。为了准确地捕捉时间序列中的重要信息，加入软注意力机制对LSTM的每一层输出进行加权求和，并与LSTM的最后一层输出拼接，作为预测层的输入。对输入的视觉/语音隐向量序列 $\boldsymbol{H}$=[H_1, H_2, …, H_N]进行如下计算：

$$\boldsymbol{h}_i = \mathrm{LSTM}_{\rightarrow}(\boldsymbol{H}_i) \oplus \mathrm{LSTM}_{\leftarrow}(\boldsymbol{H}_i), \tag{4}$$

$$\boldsymbol{A}_t = \boldsymbol{h}_N \oplus \mathrm{Attn}([\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N]) 。 \tag{5}$$

其中，$\boldsymbol{h}_i \in \mathbb{R}^{2d_h}$ 为双向LSTM在 i 时刻输出的拼接向量，$\boldsymbol{A}_t \in \mathbb{R}^{4d_h}$ 为输出的拼接向量，d_h 为LSTM的隐向量维度。

模型的预测层为一个多层感知机，A_t 经过计算，得到预测的情感得分。多层感知机由3个线性层组成，两次线性变化之间会经过一次激活函数计算，实验中使用 ReLU 激活函数。单模态情感识别任务的损失值计算方法见式(1)，$\mathcal{L}_\mathrm{v}$ 和 $\mathcal{L}_\mathrm{a}$ 分别表示视觉和声学的情感识别任务损失。在训练过程中，不对损失值进行求和，而是分别进行训练。

3 实验与结果分析

3.1 数据集

实验数据选用卡内基–梅隆大学Zadeh等发布的MOSI数据集[20]和MOSEI数据集[21]。MOSI数据集是于2016年发布的多模态情感分析数据集，包含2198条视频片段，视频内容为YouTube上的单镜头评论录像，还包含每条短视频录制者说话内容的文本。MOSEI是2018年发布的大规模情感及情绪分析数据集，内容同样来自YouTube，包含22856条视频片段。MOSI和MOSEI数据集的每条视频片段都包含一个位于[–3, 3]区间的情感得分，数值越大，正面情感极性越强。两个数据集的文本被映射为GloVe[22]词向量序列，每个词向量的尺寸为300。使用Facet面部分析工具[23]，从视频画面提取一组特征，包括面部标记、面部动作单元、头部姿势、视线轨迹和HOG特征等，从MOSI提取的每一帧的特征向量尺寸为47，MOSEI为35。使用COVAREP声学分析工具[24]，从语音提取包括12个梅尔倒谱系数(MFCCs)、音高跟踪和浊音/清音分割特征、声门源参数、峰值斜率参数和最大色散商等在内的低级的声学特征，每一帧的特征向量尺寸为74。表1列出两个数据集的详细统计数据。

由于BERT使用字节对编码[25](byte pair encoder, BPE)的分词方法，在进行模态对齐时，需要对被拆分的单词重新进行模态对齐。对拆分后多出来的token，我们使用填充0(zero)、复制(copy)和复制后平均(mean)3种方法来补充其对应的视觉和语音模态数据。图3展示文本“[CLS] no no he##s un ##fu ##nn ##y not funny at all [SEP]”分别用3种方法对齐后的形式。经过对比实验后取复制后，平均(mean)的方法。

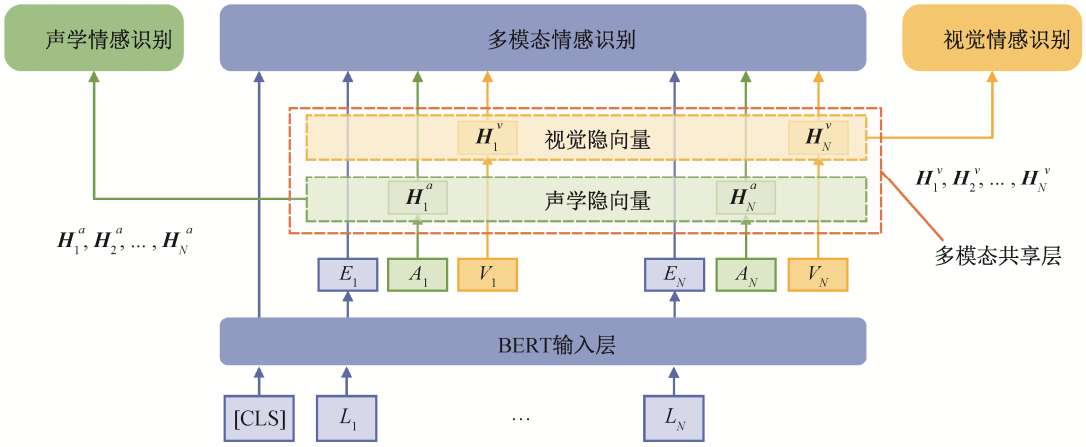


图 1 基于多任务学习的多模态情感识别框架

Fig. 1 Framework of multimodal sentiment recognition based on multitask learning

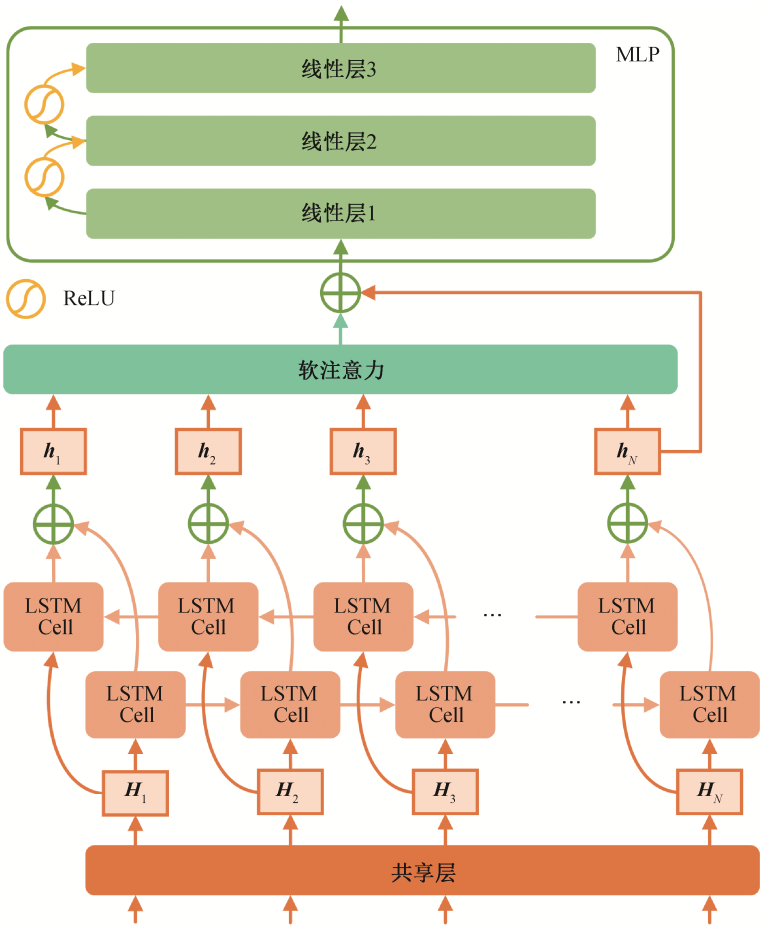


图 2 单模态情感识别模型

Fig. 2 Single-modal sentiment recognition model

3.2 训练策略及评价指标

在训练过程中，多模态模型基于 BERT 进行微调，与单模态模型一起进行训练。对多模态情感识别任务和两个单模态情感识别任务，本文都采用平均绝对误差作为损失函数，并使用 Adam 优化器[26]对模型进行参数优化。根据 Zadeh 等[20−21]的研究，

表 1　MOSI 和 MOSEI 数据集的统计信息

Table 1　Statistics of MOSI and MOSEI

数据集	视频片段数*	数据划分(训练/验证/测试)	视频片段平均长度/s	视频文本平均长度/词	最小文本长度	最大文本长度
MOSI	2198	1283/229/686	4.20	12.6	1	114
MOSEI	22856	16326/1871/4659	7.28	21.2	1	310

注: *表示已删去无法进行模态对齐的数据。

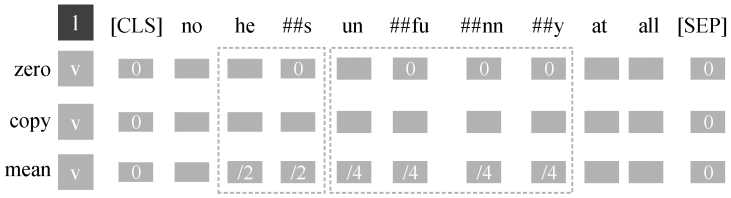


图 3　3 种模态填充方式

Fig. 3　Three modal filling methods

选取二类准确率(binary accuracy, A^2)、加权平均的 F1 值(weighted average F1-score, w-f1)、平均绝对误差(mean absolute error, MAE)和皮尔逊相关系数(Pearson correlation coefficient, Corr)作为性能评价指标。

3.3 基线模型

将本文提出的基于多任务学习的多模态情感分类模型, 与一些经典的方法和目前性能最佳(state of the art, SOTA)的方法进行对比, 以便验证其效果。

EF-LSTM[3]: 早期融合的 LSTM 模型(early-fusion LSTM)。在编码前期, 将 3 个模态{l, v, a}的特征向量进行拼接, 作为 LSTM 的输入。

LF-LSTM[3]: 晚期融合的 LSTM 模型(late-fusion LSTM)。为每个模态的特征向量分别设置一个 LSTM 网络, 用于单模态的编码, 并将 3 个 LSTM 最后一层的隐层向量进行拼接, 作为多模态的特征表示。

TFN[1]: 张量融合网络(tensor fusion network)。使用 3 个子网络分别对{l, v, a}的特征向量进行编码, 得到 z^{l},z^{v},z^{a} 3 个向量, 将 $\{z^{\mathrm{l}},z^{\mathrm{v}},z^{\mathrm{a}}\}$ 的向量尾部分别拓展一个 1, 进行外积运算, 得到融合单模态、双模态和三模态的多模态表示向量。

LMF[27]: 低秩多模态融合网络(low-rank multimodal fusion network), 是在 TFN 基础上提出的改进模型, 使用张量分解的方法分解外积运算层的参数张量。

MARN[3]: 多级注意力循环网络(multi-attention recurrent network)。基于模态间的关联是不唯一的这一观点, 采用多级注意力机制捕捉模态间的多种交互信息。

MFN[2]: 记忆融合网络(memory fusion network)。考虑 LSTM 中多个相邻时刻的信息之间的关联性, 使用跨时刻的注意力机制, 同时捕捉时序上和模态间的交互。

MTL[4]: 一种将情感识别任务和情绪识别任务联合训练的多任务学习方法。

MulT[28]: 多模态 transformer 模型(multimodal transformer)。在不改变 Transformer 编码器结构的基础上, 对其稍加改动, 提出跨模态 Transformer 网络, 实现一种模态向另一种模态的信息对齐。

M-BERT[11](SOTA): 在文本序列预训练模型 BERT 的基础上, 对其进行改造, 在 BERT 的输入端加入多模态偏移门限单元, 利用视频和语音模态信息, 使词向量在特征空间上向更能表达情感极性的方向偏移。

3.4 实验结果

表 2 为多任务学习方法和单任务学习方法在 MOSI 和 MOSEI 数据集上的评价指标实验结果。可以发现, 在 MOSI 数据集上, 多任务模型在分类指标和回归指标上都超过当前的最佳模型 M-BERT, 其中准确率提升 0.8%, 达到当前已知的最好结果。在两个回归指标上, 多任务模型也较 M-BERT 有所提升。由于 M-BERT 原论文未给出在 MOSEI 数据集上的结果, 所以表 2 中数据是我们复现的结果。在 MOSEI 数据集上, 多任务模型取得最好的分类结果, 准确率和 F1 值比 M-BERT 分别提升 1.7%和

表 2 各模型在 MOSI 测试集和 MOSEI 测试集上的结果

Table 2 Results of models on MOSI and MOSEI test sets

测试集	模型	A^2/%	w-f1/%	MAE	Corr
MOSI	EF-LSTM	72.3	72.6	1.111	0.558
	LF-LSTM	73.3	73.6	1.096	0.600
	TFN	73.9	73.4	0.970	0.633
	LMF	76.4	75.7	0.912	0.668
	MARN	77.1	77.0	0.968	0.625
	MFN	77.4	77.3	0.965	0.632
	MTL	–	–	–	–
	MulT	81.1	81.0	0.889	0.686
	M-BERT	84.4	**86.3***	0.732	0.790
	本文	**85.1**	85.2	**0.708**	**0.793**
	ΔSOTA	**↑0.8%**	–	**↓3.3%**	**↑0.4%**
MOSEI	EF-LSTM	71.4	72.1	0.720	0.497
	LF-LSTM	72.7	73.5	0.716	0.517
	TFN	74.3	73.4	0.715	0.530
	LMF	73.4	73.7	0.716	0.523
	MARN	74.3	74.7	0.712	0.533
	MFN	72.0	72.5	0.714	0.519
	MTL	80.5	78.8	–	–
	MulT	81.6	81.6	0.591	0.645
	M-BERT	82.2	82.8	**0.543**	**0.764**
	本文	**83.6**	**83.8**	0.546	0.762
	ΔSOTA	**↑1.7%**	**↑1.2%**	↑0.6%	↓0.3%

说明: *表示 M-BERT 作者未给出加权平均的 F1 值, 故用标准 F1 值代替; ΔSOTA 表示我们的方法与最佳模型在各指标上的相对变化值; ↑表示提升, ↓表示下降; 粗体数字表示效果最优, 下同。

1.2%。在回归指标上, 取得与单任务训练的 M-BERT 模型可比较的结果。

从表 2 可以看出, 多任务学习模型在两个数据集的分类指标上都取得当前最好效果, 说明引入的单模态情感识别任务可以更好地学习到具有情感倾向的视频/语音表示。在回归指标上, 多任务模型比单任务模型在小规模语料上的 MAE 提升明显, 在大规模语料上有微弱的下降。

为了确定两个辅助任务对多模态情感识别任务拟合效果的影响, 分别绘制在两个数据集的训练过程中 3 种任务的损失值曲线(图 4)。可以看到, 在 MOSI 数据集上, 两个单模态情感识别任务的损失值都能较好地拟合, 在 MOSEI 数据集上则较难拟合, 且需要更多轮的训练, 损失值才有所下降。由此可见, 在小数据集上, 加入的辅助任务能够提高多模态情感识别的拟合效果, 但在更大的数据集上, 受限于单模态编码模型的编码能力, 辅助任务难以在提高数据拟合效果上对主任务有所帮助。

表 3 展示一组样本案例。在 1 号样本中, 文本 “Maybe only 5 jokes made me laugh”包含正面情感短语“made me laugh”, 但“maybe only”又给人感觉难以确定, 单从文本很难正确地判断其中表达的情感倾向。如果只看视频内容, 能够从人物飘忽不定的眼神和紧皱的眉头判断此时带有的是负面情感, 从声学信号也可以看出人物此时的情绪并不积极, 整体的语音语调都偏低, 所以可以判断是负样本。在单任务模型上, 该样本被错误地判定为积极情绪样本, 在多任务模型上则判断正确(消极情绪), 这说明加入的两个单模态情感识别任务确实能够更好地学习到具有情感倾向的视频和语音表示。

4 消融实验

为了探究不同的共享层设置对多任务学习模型训练效果的影响, 我们在 MOSI 数据集上进行两组

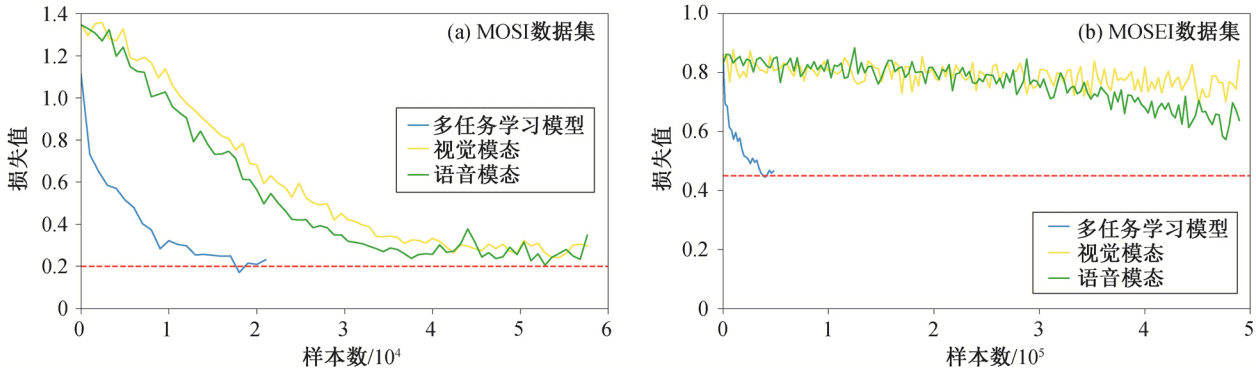
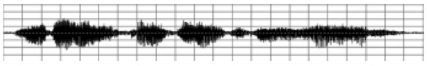
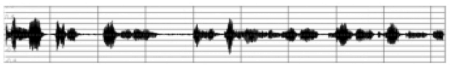
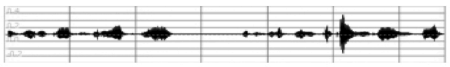


图 4　3 个任务在 MOSI 和 MOSEI 训练集上的损失曲线

Fig. 4　Loss curves of three tasks on MOSI and MOSEI train sets

表 3　多任务模型正确识别的样本案例

Table 3　Sample cases that the multitask model correctly identifies

样本		数据	情感极性	M-BERT	本文
1	文本	Maybe only 5 jokes made me laugh	Neg (−1.8)	Pos (0.4)	Neg (−0.3)
	视觉模态				
	语音模态				
2	文本	Like enough that it actually is a good reason why they why theyre attacking	Pos (1.3)	Neg (−0.2)	Pos (0.2)
	视觉模态				
	语音模态				
3	文本	I really did enjoy as well wasnt too fond of the ending	Pos (1.4)	Neg (−0.2)	Pos (0.4)
	视觉模态				
	语音模态				

消融实验。1) MM-BERT-RNN 模型。取消线性共享层，将单模态情感识别模型编码层的双向 LSTM 网络作为共享层，取 LSTM 模型的最后一层输出作为多模态情感识别模型的模态融合层的视频/语音模态输入。2) MM-BERT-RAC 模型。在 MM-BERT-RNN 模型的基础上，将单模态情感识别模型的注意力机制并入共享层，即原本的编码层成为改进后的共享层。

表 4 展示采用不同共享层策略的模型在 MOSI 数据集和 MOSEI 数据集上的评价指标实验结果，可以看到，共享层结构复杂的模型在两个数据集上的分类效果都有所下降。从回归指标看，在小数据集上，共享层结构越复杂的模型拟合效果越好，在大数据集上则相差不大。

5　结语

在多模态情感识别任务中，神经网络模型对单模态特征进行编码时，可能学习到许多与情感识别无关的特征表示。为了使模型能够学习到更具有情感倾向性的单模态表示，本文提出一种多任务多模态情感识别模型，引入视觉和声学的单模态情感识别任务共同训练，在输入端连接一个共享层，共享

表 4 不同共享层策略的模型在 MOSI 测试集和 MOSEI 测试集上的结果

Table 4 Results of models with different sharing layers on MOSI and MOSEI test sets

测试集	共享层策略	A^2/%	w-f1/%	MAE	Corr
MOSI	MM-BERT-Linear	**85.1**	**85.2**	0.708	0.793
	MM-BERT-RNN	83.5	83.6	0.700	0.801
	MM-BERT-RAC	84.3	84.2	**0.692**	**0.806**
MOSEI	MM-BERT-Linear	**83.6**	**83.8**	0.546	0.762
	MM-BERT-RNN	83.2	83.4	**0.544**	**0.763**
	MM-BERT-RAC	83.3	83.6	0.546	0.763

层后连接 3 个任务独有的模型结构。在训练的过程中, 单模态情感识别任务的作用主要是对共享层参数进行调节, 使其能够更好地捕捉对应模态的具有情感倾向的特征。实验结果表明, 我们提出的模型在 MOSI 和 MOSEI 数据集上的情感分类指标都取得当前最好的效果。

参考文献

[1] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 1103–1114

[2] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. Palo Alto, 2018: 5634–5641

[3] Zadeh A, Liang P P, Poria S, et al. Multi-attention recurrent network for human communication comprehension // Proceedings of the 32th AAAI Conference on Artificial Intelligence. Palo Alto, 2018: 5642–5649

[4] Akhtar M S, Chauhan D S, Ghosal D, et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis // Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019, 370–379

[5] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423–443

[6] Zhang C, Yang Z, He X, et al. Multimodal Intelligence: representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478–493

[7] Snoek C G M, Worring M, Smeulders A W M. Early versus late fusion in semantic video analysis // Proceedings of the 13th Annual ACM International Conference on Multimedia. New York, 2005: 399–402

[8] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: metaphor identification with visual features // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, 2016: 160–170

[9] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion // Proceedings of Structural, Syntactic, and Statistical Pattern Recognition. New York, 2014: 153–162

[10] Evangelopoulos G, Zlatintsi A, Potamianos A, et al. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia, 2013, 15(7): 1553–1568

[11] Rahman W, Hasan M K, Zadeh A, et al. M-BERT: injecting multimodal information in the BERT structure // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, 2020: 2359–2369

[12] Wang Y, Shen Y, Liu Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors // Proceedings of the 33th AAAI Conference on Artificial Intelligence. Palo Alto, 2019, 33: 7216–7223

[13] Baxter J. A model of inductive bias learning. Journal of Artificial Intelligence Research, 2000, 12(1): 149–198

[14] Thrun S. Is learning the *n*-th thing any easier than

learning the first? // Proceedings of the 8th International Conference on Neural Information Processing Systems. Cambridge MA, 1995: 640–646

[15] Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41–75

[16] Caruana R. Multitask learning: a knowledge based source of inductive bias // Proceedings of the 10th International Conference on Machine Learning. San Francisco, 1993: 41–48

[17] Duong L, Cohn T, Bird S, et al. Low resource dependency parsing: cross-lingual parameter sharing in a neural network parser // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, 2015: 845–850

[18] Sun T, Shao Y, Li X, et al. Learning sparse sharing architectures for multiple tasks // Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, 2020: 8936–8943

[19] Yang Y, Hospedales T M. Trace norm regularised deep multi-task learning [EB/OL]. (2017–02–17)[2020–09–18]. https://arxiv.org/abs/1606.04038

[20] 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

[21] Zadeh A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 2236–2246

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

[23] Zhu Q, Yeh M C, Cheng K T, et al. Fast human detection using a cascade of histograms of oriented gradients // 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, 2006: 1491–1498

[24] Degottex G, Kane J, Drugman T, et al. COVAREP — a collaborative voice analysis repository for speech technologies // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, 2014: 960–964

[25] Shibata Y, Kida T, Fukamachi S, et al. Byte pair encoding: a text compression scheme that accelerates pattern matching [R]. Technical Report DOI-TR-161. Fukuoka, 1999

[26] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2014–12–22) [2017–01–30]. https://arxiv.org/abs/1412.6980

[27] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 2247–2256

[28] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences // Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, 2019: 6558–6569