基于上下文增强 LSTM 的多模态情感分析

刘启元 张 栋 吴良庆 李寿山
(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 近年来,多模态情感分析成为了越来越受欢迎的热门领域,它将传统的基于文本的情感分析扩展到文本、图像以及声音相结合的多模态分析层面。多模态情感分析通常需要获取单模态内部的信息以及多模态之间的交互信息。为了利用每个模态中语言表达的上下文来帮助获取这两种信息,文中提出了一种基于上下文增强 LSTM 的多模态情感分析方法。具体而言,首先对于多模态的每种表达,结合上下文特征,分别使用 LSTM 进行编码,再分别捕捉单模态内部的信息;接着融合这些单模态的独立信息,再使用 LSTM 获得多模态间的交互信息,从而形成多模态特征表示;最后采用最大池化策略,对多模态表示进行降维,从而构建情感分类器。该方法在 MOSI 数据集上的 *ACC* 值达到 75.3%,*F*1 达到了 74.9。相比传统的机器学习方法(如 SVM),所提方法的 *ACC* 值高出 8.1%,*F*1 值高出 7.3。相比目前较为先进的深度学习方法值,所提方法在 ACC 值上高出 0.9%,*F*1 值上高出 1.3,与此同时可训练参数量只有之前方法的 1/20,训练速度提高了约 10 倍。大量的对比实验结果表明,相比传统的多模态情感分类方法,所提方法的性能有显著提升。

关键词 多模态,情感分析,上下文增强
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.181001941

Multi-modal Sentiment Analysis with Context-augmented LSTM

LIU Qi-yuan ZHANG Dong WU Liang-qing LI Shou-shan
(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract In recent years, multi-modal sentiment analysis has become an increasingly popular research area, which extends traditional text-based sentiment analysis to a multi-modal level that combines text, images and sound. Multi-modal sentiment analysis usually requires the acquisition of independent information within a single modality and interactive information between different modalities. In order to use the context information of language expression in each modality to obtain these two kinds of information, a multi-modal sentiment analysis approach based on context-augmented LSTM was proposed. Specifically, each modality is encoded in combination with the context feature using LSTM which aims to capture the independent information within single modality firstly. Subsequently, the independent information of multi-modality is merged, and the other LSTM layer is utilized to obtain the interactive information between the different modalities to form a multi-modal feature representation. Finally, the max-pooling strategy is used to reduce the dimension of the multi-modal representation, which will be fed to the sentiment classifier. The method achieves 75.3% ACC on the MOSI data set and F1 reaches 74.9. Compared to traditional machine learning methods such as SVM, ACC is 8.1% higher and F1 is 7.3 higher. Compared with the current advanced deep learning method, it is 0.9% higher on ACC and 1.3 higher on F1. At the same time, the trainable parameters are reduced by about 20 times, and the training speed is increased by 10 times. The experimental results demonstrate that the performance of the proposed approach significantly outperforms the competitive multi-modal sentiment classification baselines.

Keywords Multi-modal, Sentiment analysis, Context enhancement

1 引言

多模态研究是一个越来越受重视的研究领域。传统的自然语言处理任务一般只使用单一的文本信息,而多模态任务则使用文本、图像、语音等多种模态信息。常见的多模态任务包括多模态情感分析[1]、情绪识别[2]和人格特质识别[3]等。

对于多模态情感分析任务,其核心挑战在于如何更好地利用模态内部信息(intra-modality)和模态之间的交互作用

到稿日期:2018-10-18 返修日期:2019-04-01 本文受国家自然科学基金(61331011,61375073)资助。
刘启元(1994-),男,硕士生,CCF 会员,主要研究方向为自然语言处理、情感分析,E-mail:qyliu@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;**吴良庆**(1995-),男,硕士生,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,E-mail:lishoushan@suda.edu.cn(通信作者)。

(inter-modality)信息。模态内部信息就是单个独立的模态所能被挖掘并利用的信息;而模态之间的交互作用则是不同模态之间的相互关联与联系所能带来的有用信息[4-5]。如何利用不同模态之间的交互信息,也是多模态任务与单模态任务的最大区别。

模态内部信息通常由卷积网络和长短期记忆网络捕获[6],而模态之间的交互作用则有多种不同的方法。例如早融合(Early Fusion)[7],这种方法会在信息输入时直接将多种模态信息进行拼接操作。与之对应的是晚融合(Late Fusion)[8],晚融合则是把每种模态信息单独进行训练,考虑模态内在的信息,然后执行决策投票。之后也提出了张量融合(Tensor Fusion)[8]这种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的方法,它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。可见,若要提高多模态情感分析任务的准确率,不仅要深度挖掘模态的内部信息,更要处理好模态之间的交互作用。

以上传统方法大都忽略了每个表达的上下文信息。上下文信息加上多模态信息对于情感分析任务具有明显的帮助作用,如图 1 所示。

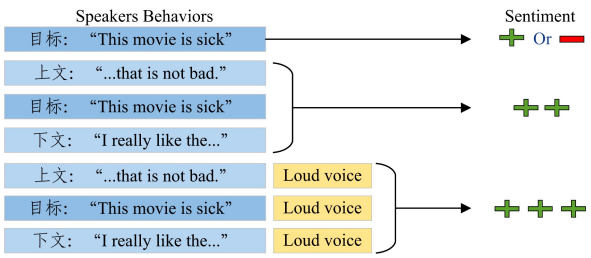


图 1 多模态及上下文增强对情感分析的影响例子

Fig. 1 Example of effects of multi-modality and context enhancement on sentiment analysis

图 1 中,如果只有单独的文本信息"This movie is sick",那么这句话可以被理解为消极或者积极。但如果加入上文部分信息"…that is not bad."和下文部分信息"I really like the …",那么这句话就很容易被理解为积极的情感,如果再配上高亢的声音信息,那么就更增加了这句话所传递的正向信息。

虽然 Poria 等[5]曾经尝试引入上下文信息来帮助情感分析,但是所提出的是一种分步的过程,先学习模态内部信息,训练完模型后获得中间隐层表示,再构建模型学习交互信息。该方法没有把模态内部信息和交互信息结合在一起进行联合学习。

为了解决上述问题,本文提出了一种基于上下文增强 LSTM 的多模态情感分析方法。该方法的特色在于利用多模态的上下文信息,层次化地使用 LSTM 网络来捕获单模态的独立信息以及多模态间的交互信息。具体而言,首先,结合上下文特征对多模态的每个表达分别使用 LSTM 进行编码,并分别捕捉单模态内部的信息;其次,融合这些单模态内部的独立信息,再经过一层 LSTM 即可获得多模态间的交互信息,从而形成多模态特征表示;最后,采用最大池化策略来对多模态表示进行降维,从而构建情感分类器。

为了验证所提方法的有效性,采用文本和音频两个模态进行口语情感分类实验。实验结果表明,相比传统的多模态情感分类方法,所提方法在多模态情感分类上的性能提升得十分明显。

2 相关工作

尽管有大量针对声音和视觉的研究,但是将其与文本相结合的研究相对较少。早期的工作运用特征融合的手段将词汇和手工标注的低级声音特征融合在一起[9]。而近年来,深度学习被用来提取更高级别的多模态特征。双向 LSTM 被用来从声学描述和视觉特征中学习长期依赖性[10-11]。CNN 可以将提取的文本[4]和视觉特征[6]用于特征融合的多核学习。Wang 等提出了 SCL-CNN 模型,并将该模型用于对文本与声音相结合的情感分类任务进行探索[12]。Gu 等[13]使用深度神经网络进行特征级别的融合,Zadeh 等提出的张量融合网络[8]和图记忆融合网络[14]都取得了不错的效果,但是这些方法在时间和空间上的复杂度较高,尤其是后者,还需依赖动态记忆网络方法。还有一些工作进行了单词级别的融合[15],Poria 等提出了一种基于 LSTM 的模型,该模型能捕捉来自同一视频的上下文信息,而捕捉模态的内部信息和模态间的交互信息是一个分步的过程。

Zadeh 等于 2016 年发布了数据集 MOSI[16],该数据集从 93 个意见表达视频中抽取了 2 199 个观点段,视频的主要内容是视频作者面对摄像头来表达自己对某一事物的观点。可以看出,每个视频作者都有若干个观点段与之对应。同一作者的观点段之间可能存在某种联系,可以利用这些联系来进行情感分析。

基于以上相关研究工作的讨论,我们提出了基于上下文增强的文本、语音融合的情感分析方法。不同于已有研究,我们的模型和方法主要有以下特点:1)采用了全新的上下文增强方式,能够捕捉同一视频内的多模态上下文信息;2)不同于以往利用上下文信息方法使用的分步过程,本文方法联合学习模态的内部信息和模态间的交互信息;3)本文模型的训练参数更少,训练速度更快。

3 基于上下文增强 LSTM 的多模态情感分析方法

图 2 为本文提出的基于上下文增强 LSTM 的多模态情感分析模型框架图,简称 CAL 模型。该模型主要分为以下 5 个部分:

(1)上下文特征表示。该部分的主要任务是将对齐后的文本和声音特征经过上下文信息增强后输入到神经网络。

(2)单模态独立信息。这一部分用来挖掘单独模态的内部信息(intra-modality)。

(3)多模态交互信息。这一层是将上一步得到的各个模态的内部信息融合起来,从而得到模态之间的交互作用(inter-modality)。

(4)特征降维。该层主要使用了 Max pooling 的方法来得到最有用的多模态信息。

(5)情感分类。运用之前所得到的多模态融合信息进行情感分类。

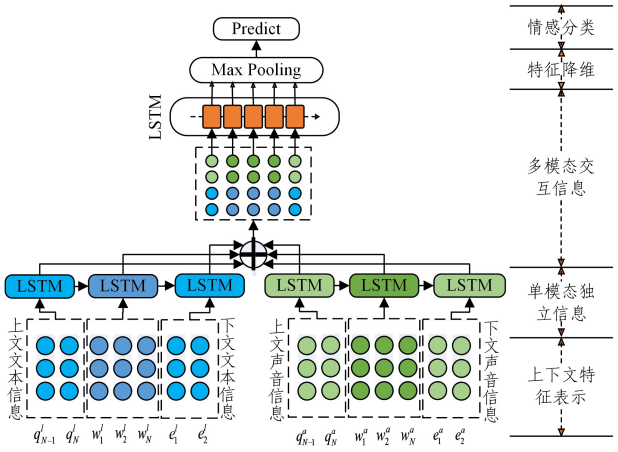


图 2　基于上下文增强 LSTM 的多模态情感分析模型框架图

Fig. 2　Framework of multi-modal sentiment analysis model based on context-enhanced LSTM

下文将对以上 5 个部分进行详细介绍。

3.1　上下文特征表示

我们采用了由卡内基梅隆大学提供的多模态数据 SDK[1)] 来获取文本、语音的特征。文本特征为 GloVe 词向量[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得到的[18]。因为不同模态之间的信息需要严格对齐才能获取最好的性能[19]，所以采用了 SDK 中的 P2FA 方法将每个单词时间段内的声音信息融合后再与单词对齐。这样的对齐方式更加直观，并且符合人类的表达方式。

经过以上方法得到了可以输入神经网络训练的文本和声音特征。假设一个句子包含 n 个单词，那么对齐后的声音特征也会有 n 个，我们用 w_t^m 来表示文本或者声音中的第 t 个特征，其中 $m=\{l,a\}$，l 代表文本信息，a 代表声音信息。用 q_t^m 表示上文的第 t 个文本或声音特征，用 e_t^m 表示下文的第 t 个文本或声音特征。那么我们的输入可以表示为：

$$W^m=[w_1^m;w_2^m;\cdots;w_n^m] \tag{1}$$

$$Q^m=[q_{n-r+1}^m;q_{n-r+2}^m;\cdots;q_n^m] \tag{2}$$

$$E^m=[e_1^m;e_2^m;\cdots;e_r^m] \tag{3}$$

其中，r 表示上下文特征的增强强度，$W^m\in\mathbb{R}^{n\times d^m}$ 表示模态 m 的信息；$Q^m\in\mathbb{R}^{r\times d^m}$ 表示模态 m 的上下文信息；$E^m\in\mathbb{R}^{r\times d^m}$ 表示模态 m 的下文信息。

3.2　单模态独立信息

这一部分主要用来挖掘单独模态的内部信息，主要方法是采用各个模态私有的 LSTM 来获取单个模态的语义信息，为了更好地保留模态信息，我们返回了每一个时刻的隐藏表示，具体表示如下：

$$h_{x_t}^m=LSTM(x_t^m) \tag{4}$$

$$H_x^m=[h_{x_1}^m;h_{x_2}^m;\cdots;h_{x_n}^m],H_x^m\in\mathbb{R}^{n\times d^m} \tag{5}$$

$$H^m=H_q\oplus H_w\oplus H_e,H^m\in\mathbb{R}^{(2r+n)\times d^m} \tag{6}$$

其中，$x=\{q,w,e\}$，q,w,e 分别对应式(1)、式(2)、式(3)，分别表示上文信息、目标句子和下文信息。式(4)中的 $h_t^{m,x}$ 是 LSTM 在第 t 时间步的隐层输出。值得注意的是，LSTM 之间是有信息流动的，如编码上文信息的 LSTM 层的最后一步输出会作为编码主文信息 LSTM 层的初始状态。式(5)中的 H_x^m 是经过 LSTM 后所有时间步的隐层表示。于是，通过式(6)的融合操作，我们就能分别得到经过上下文增强后的文本表示 H^l 和声音表示 H^a。

3.3　多模态交互信息

在 3.2 节中分别得到了文本和声音的 LSTM 隐层表示，而本节的主要任务是对两个模态的信息进行融合，并挖掘出模态之间的交互作用。首先对两个模态的隐层表示进行简单拼接，该步骤用到的公式如下：

$$H^{l,a}=H^l\oplus H^a,H^{l,a}\in\mathbb{R}^{(2r+n)\times(d^l+d^a)} \tag{7}$$

根据式(7)对两个模态的信息进行最内层维度的拼接，从而得到文本和声音简单的融合信息 $H^{l,a}$，但是该融合信息并不能反映两个模态间的交互作用，因此又对信息进行了深度融合：

$$M_h^{l,a}=LSTM(H^{l,a}) \tag{8}$$

$$m^f=\tanh(W\cdot m_h^{l,a}+b) \tag{9}$$

首先，通过设置一个多模态的上下文增强 LSTM 用于捕捉交互特征，将之前得到的简单融合特征 $H^{l,a}$ 放入该层，从而得到了每一个时间步的隐藏表示，将其记为 $M_h^{l,a}$。之后，将 $M_h^{l,a}$ 的其中一行特征记为 $m_h^{l,a}$。式(9)中，m^f 是 $m_h^{l,a}$ 通过 tanh 层后的表示，W 和 b 分别是 tanh 层的权重和偏置。最后，用 $M^f=\lceil m_1^f;m_2^f;\cdots;m_{2r+n}^f\rceil$ 表示文本和声音深度融合后的特征，但此时我们还不能直接用此特征进行分类，还要对特征进行筛选。

3.4　融合特征降维与情感分类

在 3.3 节中，我们得到了两个模态的融合信息，本节将对该信息进行筛选。筛选的过程如下所示：

$$M^F=\max\ pooling\{M^f\} \tag{10}$$

此公式最大池化层的具体操作是对 M^f 的每一列取最大值并将其保留，值最大代表其在该列特征中所表示的特征最强，更有利于接下来的分类。

在得到经过筛选的融合特征后，使用 tanh 和 sigmoid 进行最终的分类预测，具体的表示如下：

$$\hat{y}=\sigma(W_t\cdot(\tanh(W_s\cdot M^F+b_s))+b_t) \tag{11}$$

其中，$\hat{y}$ 就是最终的情感分类结果，W_t,W_s,b_t,b_s 分别对应 tanh 和 sigmoid 层的权重和偏执。

3.5　优化策略

在模型训练过程中，我们选取了交叉熵误差作为损失函数，损失函数的计算公式如下：

$$Loss(\hat{y},y)=-\sum_{s=1}^{S}\sum_{c=1}^{C}y_s^c\cdot\log\hat{y}_s^c \tag{12}$$

其中，y 是真实标签，$\hat{y}$ 是模型预测的概率，S 是训练样本总数，C 是类别的数目。这里的情感分类实际上是一个二分类问题。同时，实验中采取 adam 优化器来优化模型参数[20]。

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

4 实验

本节将系统分析本文提出的方法在多模态情感分析任务上的效果。

4.1 实验设置

本文所做的实验都是基于卡梅隆大学所提供的 MOSI(Multimodal Opinion-level Sentiment Intensity)数据集的,主要工作是对该数据集进行情感分类。MOSI 数据集中包含了从 YouTube 上找到的 93 个意见表达视频,并从中抽取了 2199 个含有情感的视频段,最后抽取每个视频段中的文本、图片以及声音特征,并对其人工标注情感强度后将其作为一条多模态数据。具体的统计数据如表 1 所列。其中,训练集有 1283 个视频段,验证集有 230 个视频段,测试集有 686 个视频段(同一个视频中的数据只能出现在一个集合中)。由于每条数据都标注了其所在的视频 ID 和段落 ID,因此可以判断不同数据之间是否存在上下文关系。我们将情感强度大于 0 的标签记为正类,即该条数据表达了积极情感;将情感强度小于 0 的记为负类,即该条数据表达了消极情感。

表 1 MOSI 数据集的统计数据

Table 1 Statistics of MOSI dataset

描述	数量
视频总数	93
演讲者数量	93
句子总数	2199
每个视频平均句子数量	23.2
句子平均长度/s	4.2
总单词个数	26295
单词数量(不重复)	3107

对于实验结果的评估,我们使用了两种评估值,一个是准确率,记为 *ACC*,另一个是 F1 分数,记为 *F*1。

实验的所有代码都采用 Keras 深度学习架构编写,并使用了 Theano 后端。运行环境为 Linux 操作系统,并使用 Nvidia Titan XP GPU 进行加速。

4.2 实验

实验结果如表 2 所列,以下将简单描述表 2 中对应的方法。

1)SVM-MD:其是一个利用早融合训练多模态特征的 SVM 模型[21]。Morency 等[1]和 Perez-Rosas 等也在多模态级联特征上使用了 SVM。

2)RF:使用随机森林[22]作为非神经网络的分类器。

3)CNN:用于多模态情感分析的七层卷积神经网络架构[5]。

4)SAL-CNN:选择-附加学习方法,可以提高训练神经网络在多模态情感分析中的普遍性[12]。

5)CDSA:一种上下文依赖的分层次情感分析方法。

6)EF-LSTM:早融合 LSTM[23],在每个时间步连接来自不同模态的输入,并将其用作于单个 LSTM 的输入。

7)TFN:通过创建多维张量来明确模拟单个模态和交叉模态特征,该多维张量捕获 3 种模态中的单模态、双模态以及三模态的交互作用[8]。我们还原了该方法,并在文本和声音两个模态上进行了实验,最后得到了实验数据。

8)CAL:是本文提出的方法,其中 CAL(without_context)保持原有模型不变,但是不使用上下文增强,使用了上下文增强的模型。

表 2 不同模型在 MOSI 数据集上的结果

Table 2 Results of different models on MOSI datasets

方法	*ACC*/%	*F*1
SVM-MD	67.2	67.6
RF	67.4	66.8
CNN	68.7	68.4
EF-LSTM	70.6	71.2
SAL-CNN	72.5	73.6
CDSA	73.6	74.0
TFN	74.4	73.6
CAL(without_context)	74.2	74.1
CAL(with_context)	75.3	74.9

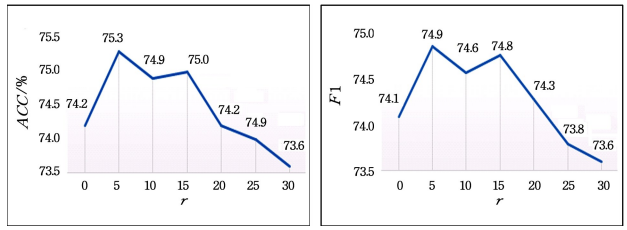
对比表 1 中的实验数据可以看出,所提方法不管是在 *ACC* 上还是在 *F*1 上都取得了最好的结果。首先,相比一些早期的方法(如 SVM 和 RF),所提方法的结果遥遥领先,*ACC* 分别高出了 8.1%和 7.9%,*F*1 分别高出了 7.3 和 8.1,即使对比近期较为先进的方法 TFN,所提方法的 *ACC* 也高出 0.9%,*F*1 高出 1.3,所提方法仍然表现出了优异性能;其次,对比我们的模型在没有加入上下文信息时的效果,加入上下文增强后的性能有了明显的提升,可见上下文增强对本实验任务确实是有明显帮助的;最后,同样是使用了加入上下文信息的思想,我们手动实现的 CDSA 的效果不如提出的 CAL,*ACC* 相差 1.7%,*F*1 相差 0.9。这些数据都证明了所提方法和模型在文本和声音上进行情感分析的优异性能。

此外,TFN 总共包含约 2×10^7 个可训练参数,而我们的 CAL 只有约 1×10^6 个可训练参数,TFN 训练参数的数量约为所提模型的 20 倍。训练速度方面,TFN 的训练平均速率为 131IPS(每秒训练数据的条数),而所提模型达到了 1273IPS,速度约为 TFN 的 10 倍。

通过对比实验可以看出,所提模型虽然参数更少,训练速度更快,但是训练效果却远远好于其他模型,由此可见,本文方法的性能相对优越。

4.3 实验结果分析

对于 CAL 模型,一个特有的参数是式(2)、式(3)中的 r 值,也即引入信息量的多少。我们用 CAL 模型在 MOSI 数据集上进行了不同信息量的对比实验,图 3(a)、图 3(b)是不同信息量对 *ACC* 和 *F*1 两个评估值的影响变化统计图,其中纵坐标是评估值的结果,横坐标是数据量,也即 r 的值。可以看到,当 $r=5$ 时,*ACC* 与 *F*1 同时达到峰值,随着 r 的增加,两个评估值反而下降。但是当 r 值适当时,上下文增强的效果是比较明显的。



(a)ACC 变化趋势 (b)F1 变化趋势

图 3 不同上下文信息量的对比实验结果

Fig. 3 Comparison test results of different context information

因此，我们可以得到一个结论：适当引入上下文信息对于结果具有明显的提高作用，但是过多的上下文增强信息对于情感分析的效果会减弱。

结束语　本文提出了一种基于上下文增强 LSTM 的多模态情感分析方法。首先，通过上下文增强的方法使得模态内部信息更加充足；其次，运用多模态特征融合和特征降维的方法捕捉模态之间的交互作用；最后，再通过分类器进行情感分类以最终完成情感分析的任务。在 MOSI 数据集的对比实验中，相较于其他方法，本文方法在文本和声音两个模态中取得了最好的效果；我们还探索了不同信息量对于上下文加强效果的影响。以上实验均证明了我们方法的有效性和优异性能。

在未来的工作中，我们将会不断完善现有方法与模型，将本文方法运用到其他数据集中。此外，我们还将继续在多模态的情感强度上进行研究与探索。

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] BUSSO C, BULUT M, LEE C C, et al. Iemocap: Interactive emotional dyadic motion capture database [J]. Journal of Language Resources and Evaluation, 2008, 42(4): 335-359.

[3] PARK S, SHIM H S, CHATTERJEE M, et al. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach [C] // Proceedings of the 16th International Conference on Multimodal Interaction. New York: ACM, 2014: 50-57.

[4] 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. Stroudsburg, PA: ACL, 2015: 2539-2544.

[5] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos [C] // Proceedings of the 55th ACL. 2017: 873-883.

[6] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis [C] // IEEE 16th ICDM. Piscataway, NJ: IEEE, 2016: 439-448.

[7] 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.

[8] 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. Stroudsburg, PA: ACL, 2017: 1103-1114.

[9] DIANE J L, KATE F R. Predicting student emotions in computer-human tutoring dialogues [C] // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2004: 351.

[10] EYBEN F, WOLLMER M, GRAVES A, et al. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues [J]. Journal on Multimodal User Interfaces, 2010, 3(1/2): 7-19.

[11] WOLLMER M, WENINGER F, KNAUP T, et al. Youtube movie reviews: Sentiment analysis in an audio-visual context [J]. IEEE Intelligent Systems, 2013, 28(3): 46-53.

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

[13] GU Y, CHEN S H, MARSIC I. Deep multimodal learning for emotion recognition on spoken language [C] // 2018 IEEE International Conference Proceedings of Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2018.

[14] ZADEH A, LANG 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. Stroudsburg, PA: ACL, 2018: 2236-2246.

[15] 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.

[16] ZADEH A, ZELLERS R, PINCUS E, et al. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos [J]. arXiv: 1606.0659.

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

[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] 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. Menlopark, CA: AAAI, 2018.

[20] KINGMA D P, BA J. Adam: A method for stochastic optimization [J]. arXiv: 1412.6980.

[21] 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.

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

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