控制与决策
Control and Decision

基于深度情感唤醒网络的多模态情感分析与情绪识别

张峰, 李希城, 董春茹, 花强

引用本文:
张峰,李希城,董春茹,花强. 基于深度情感唤醒网络的多模态情感分析与情绪识别[J]. 控制与决策, 2022, 37(11): 2984−2992.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0782

基于深度情感唤醒网络的多模态情感分析与情绪识别

张 峰, 李希城, 董春茹, 花 强†

(1. 河北大学 数学与信息科学学院，河北 保定 071002；
2. 河北省机器学习与计算智能重点实验室，河北 保定 071002)

摘 要: 随着网络平台上各类图像、视频数据的快速增长,多模态情感分析与情绪识别已成为一个日益热门的研究领域. 相比于单模态情感分析,多模态情感分析中的模态融合是一个亟待解决的关键问题. 受到认知科学中情感唤起模型的启发,提出一种能够模拟人类处理多通道输入信息机制的深度情感唤醒网络(DEAN),该网络可实现多模态信息的有机融合,既能处理情绪的连贯性,又能避免融合机制的选择不当而带来的问题. DEAN网络主要由以下3部分组成:跨模态Transformer模块,用以模拟人类知觉分析系统的功能;多模态BiLSTM系统,用以模拟认知比较器;多模态门控模块,用以模拟情感唤起模型中的激活结构. 在多模态情感分析与情绪识别的3个经典数据集上进行的比较实验结果表明,DEAN模型在各数据集上的性能均超越了目前最先进的情感分析模型.

关键词: 多模态情感分析；多模态情绪识别；深度情感唤醒网络；跨模态Transformer；多模态BiLSTM系统；多模态门控机制

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0782

引用格式: 张峰,李希城,董春茹,等. 基于深度情感唤醒网络的多模态情感分析与情绪识别[J]. 控制与决策, 2022, 37(11): 2984-2992.

Deep emotional arousal network for multimodal sentiment analysis and emotion recognition

ZHANG Feng, LI Xi-cheng, DONG Chun-ru, HUA Qiang†

(1. College of Mathematics and Information Sciences，Hebei University，Baoding 071002，China；2. Hebei Key Laboratory of Machine Learning and Computational Intelligence，Baoding 071002，China)

Abstract: Multimodal sentiment analysis and emotion recognition have become an increasingly popular research topic, with a large number of emerging of various images and videos on internet. Compared with the unimode sentiment analysis, one of the key issues in multimodal sentiment analysis is the fusion strategy. Inspired by the emotional arousal model in cognitive science, a deep emotional arousal network (DEAN) is proposed, which can simulate emotional coherence. The proposed DEAN consists of three parts: A crossmodal transformer module that simulates the functions of a human perception analysis system, a multimodal BiLSTM system that simulates the cognitive comparator, and a multimodal gating module that is used to simulate the activation mechanism in the physiological emotional arousal model. Extensive experimental comparisons on three benchmarks for multimodal sentiment analysis and emotion recognition are conducted, and the experimental results show that the DEAN outperforms the state-of-the-art methods.

Keywords: multimodal sentiment analysis；multimodal emotion recognition；deep emotional arousal network；crossmodal Transformer；multimodal BiLSTM system；multimodal gating module

0 引 言

人类的情绪由神经元回路控制,神经回路通过生理唤醒收集情绪信息并产生情绪行为[1]. 在人类大多数的交际场景中,人们除了需要处理语言信息之外,还需要处理声音和视觉等多通道的信息. 伴随着近几年网络平台的普及应用,各类图像、视频数据的快速增长,多模态情感分析与情绪识别已成为一个日益热门的研究领域. 海量的多模态数据中蕴含着人们丰富的态度和看法,揭示其背后隐含的情感在票房预测[2]、政治选举[3]、商品推荐[4]等方面拥有较大的应

收稿日期: 2021-05-05；**录用日期:** 2021-08-09.

基金项目: 国家自然科学基金面上项目(61773150, 61672205)；河北省自然科学基金面上项目(F2018201115)；河北省教育厅科学技术研究重点项目(ZD2019021)；河北大学高层次创新人才科研启动经费项目.

责任编委: 侯忠生.

†通讯作者. E-mail: huaq@hbu.edu.cn.

用价值. 因此如何模拟人类处理多通道输入信息的过程,对多模态数据进行有机融合,提高情感分析的准确度,已成为目前多模态情感分析的主要研究问题之一[5-6].

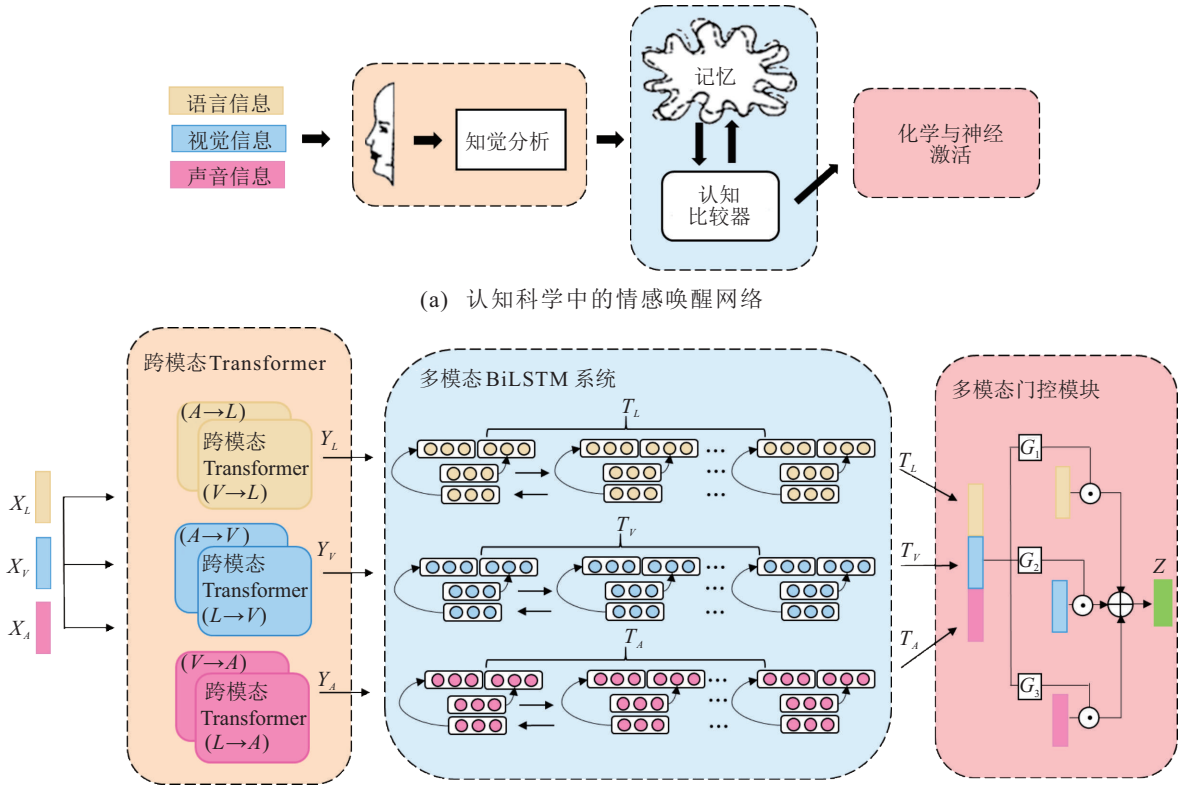


图 1　情感唤醒模型与深度情感唤醒网络

近年来,对多模态数据融合方法的研究主要有: 1) 基于显式融合方式的多模态融合,包括特征级融合、决策级融合和混合策略的融合; 2) 基于隐式融合方式的多模态融合,包括双线性融合[7]、条件随机场融合[8]、基于张量的融合[9]和基于注意力的融合[10-11]等方法. 其中基于注意力的融合方法是近年来应用最广泛的融合策略,其代表模型为基于Transformer的情感识别模型[10-13]. 该模型可利用注意力机制学习模态信息之间的依赖关系,因其结构上可以实现并行化处理而被广泛应用于多模态数据场景[10,13-14]. 然而,基于Transformer的情感识别模型仍面临两个主要挑战: 1) 基于Transformer的模型难以模拟人类情绪之间的连贯性. 由于基于Transformer的模型大多采用并行机制,因此无法很好地建模人类连续变化的情绪. 2) 以Transformer为代表的基于注意力机制的融合策略通常忽略不同模态的重要性差异[10]. 而最新的实验结果表明,不同模态对融合结果的影响是不同的,例如语言模态往往比视觉与听觉模态的影响更大[15].

本文受认知科学中的情感唤醒模型[16]的启发,提出一种深度情感唤醒网络(deep emotional arousal network)(图1(b)). DEAN从仿生的角度出发,构建跨模态Transformer模块、多模态BiLSTM系统和多模态门控模块,分别用于模拟认知科学中情感唤起模型中的知觉分析系统、认知比较器和激活机构3个亚系统(图1(a))的功能. DEAN首先采用跨模态Transformer模块来建立不同模态信息间的相互增强作用,通过跨模态Transformer模块进行空间上的跨模态交互,实现多模态信息在特定时刻的知觉分析过程; 然后,利用多模态BiLSTM系统在跨模态Transformer模块基础上引入时间依赖关系,使当前时刻的模态特征包含过去时刻的情感信息,实现对于情感连贯性的模拟; 最后,采用多模态门控模块自适应地对各目标模态的输出信息进行控制,隐式实现对不同模态信息的加权融合. 整体来说,DEAN模型模拟人类处理多通道信息的情感交流过程,为多模态情感分析与情绪识别提供了一个完整框架.

1　相关研究

多模态情感分析与情绪识别的任务是基于语言、视觉和听觉等多种模态输入信息来预测用户所表达的情感倾向或情绪. 相比于单模态情感分析任务而

言,多模态情感分析任务将情感分析任务从文本情感分析拓展到多个模态,一方面增加了任务的复杂度,同时也增加了情感识别任务的应用范围[17]. 多模态情感分析与情绪识别任务的关键挑战之一在于多模态的融合策略是否有效. 目前多模态融合策略包括显式融合和隐式融合两种方式. 显式融合的结果通常取决于融合的方式,不依赖于特定的分类方法或者回归方法;隐式融合往往需要先隐式构建一种融合模型,然后利用该模型解决多模态融合问题.

1.1 基于显式融合策略的多模态情感分析模型

显式融合是最早提出的多模态融合策略. 根据融合多模态输入信息的阶段不同,显式融合方法可分为特征级融合(早期融合)、决策级融合(后期融合)和混合融合. 早期融合策略通常需要首先学习输入信息的代表性特征,然后通过特征连接或加权组合的方式将提取的特征进行融合. 由于深度网络强大的表示能力,最近提出的早期融合方法常采用卷积神经网络(CNN)[18-20]、递归神经网络(RNN)[21]、长短期记忆网络(LSTM)[22-23]等神经网络模型进行特征提取并实现融合. 然而,由于早期融合的方法大多采用分别对不同模态信息进行特征提取的方式,难以对模态之间的相互关系有效建模. 后期融合的方法在决策的输出层级对各模态的分类结果进行融合,包括独立训练单模态分类器以及执行决策投票等[24]. 后期融合模型的结果往往优于各单一模态模型的表现,但后期融合模型无法有效地学习各模态内的动态.

混合融合的方法[25]结合早期融合和后期融合方法的优点,通过模态内和模态间两种建模方式共同作用[26]来提高情感分析和识别的精确度. 混合融合的方法相比早期融合与后期融合更多样化[9,11],因此取得了较好的效果. 这些显式融合方法在深度网络有效特征提取的帮助下,提高了情感识别的精度,但识别结果的好坏依旧无法摆脱融合方法选择的影响.

1.2 基于隐式融合策略的多模态情感分析模型

隐式融合策略,是一种基于模型的融合策略. 早期基于模型的融合方法包括多核学习[27]、双线性融合[7]和概率图模型[8]. 近几年研究者们也逐渐提出了基于深度网络构建融合模型的方法,代表性工作有:1) 基于张量的融合,如张量融合网络(TFN)[9]、低秩多模态融合(LMF)[28]和局部约束多模态融合网络(LCMFN)[29]等;2) 基于转换的融合,如模态转换模型(MCTN)[30]和Seq2Seq模态转换模型(SSMT)[31]等;3) 基于注意力的融合,如基于多注意力的循环网络(MARN)[32]、多模态Transformer (MulT)[10]、循环变化嵌入网络(RAVEN)[11]等. 其中基于注意力的融合方法是近年来被广泛应用的一种融合策略. 最新的实验结果[15]表明,与其他融合方法相比,基于注意力的融合策略较大地提高了情感分析模型的性能. 原因可能是基于注意力的融合模型可以从全局的角度模拟不同模态的交互动态与内部动态. 然而,基于注意力机制的方法因其复杂的模型结构和大量的参数,比其他的模态融合方法需要更多的训练时间,故常采用并行结构. 并行结构往往无法很好地建模人类连续变化的情绪.

综上,现有的多模态情感分析模型在很大程度上改进了情感分析任务的精度,但没有形成一个完整的处理多模态输入信息的通信框架. 而最先进的基于Transformer的并行结构又忽视了人类情绪的连贯性,无法区分不同模态对于融合结果的影响. 因此,受认知科学的情感唤起模型启发,提出一种深度情感唤醒网络(DEAN). DEAN以多模态信息为输入,提供一个全面的模拟人类情感交流的模型,避免了多模态输入信息融合方式选择的困扰. 该模型的优点主要包括:1) 模拟认知科学中情感唤起模型,为多模态的情感分析问题提供一个完整的解决框架,避免了对多模态输入信息融合方式的选择;2) 在Transformer并行结构中引入时间维度,实现对情感连续性进行建模,用于描述人类情绪的连贯性;3) 通过嵌入多模态门控模块区分不同模态的重要性. 在CMU-MOSI、CMU-MOSEI、IEMOCAP数据集上进行的实验结果表明,DEAN的识别精度均超过了目前最先进模型的性能.

2 深度情感唤醒网络

本节将根据认知科学中的情感唤醒模型提出深度情感唤醒网络. DEAN由以下3部分组成.

1) 跨模态Transformer模块,该模块模拟认知科学中的情感唤醒模型的知觉分析系统. 具体地说,它利用注意力机制实现了模态间信息在空间上的融合,利用辅助模态增强了目标模态特征.

2) 多模态LSTM系统,该结构用于模拟生理学上情感唤醒模型的认知比较器. 此过程可由3个步骤完成:

① 利用BiLSTM网络从各目标模态输入特征;

② 将当前提取的特征与过去的记忆进行比较;

③ 在时间序列上实现信息传递,模仿人类情绪的连贯性.

3) 多模态门控模块,采用多模态门控模块来模拟生理学上情感唤醒模型的激活结构. 该模块可以根据目标模态的重要性控制目标模态的输出信息,隐式地融合来自不同模态的信息.

2.1　跨模态Transformer模块

DEAN采用跨模态Transformer模块[10]对神经系统的多通道感知分析过程进行建模. 本文主要考虑3种输入模态,即语言、视觉和音频,通过3对跨模态Transformer的前馈融合过程隐式融合多模态输入. 每个跨模态Transformer利用注意力机制在低层特征级别上增强目标模态特征. 多模态输入信息中的任意一种模态均可以作为目标模态,其余两种则视为辅助模态. 例如将语言模态(L)设置为目标模式,则以语言模态为目标的跨模态融合包括视觉对语言的融合$V \to L$和音频对于语言的融合$A \to L$.

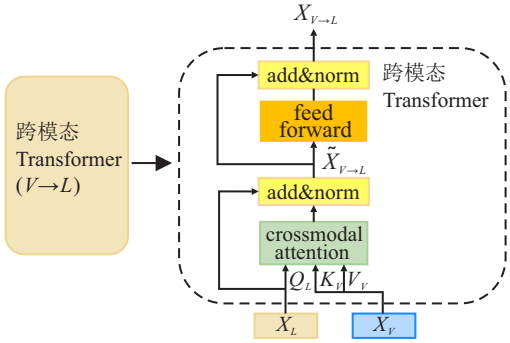


图 2　跨模态Transformer模块

以跨模态Transformer $V \to L$为例(图2),跨模态间的信息交互过程可描述如下: 首先,将语言模态与视觉模态输入信息通过Embedding转化为特征向量$X_L \in R^{T\times d_L}$和$X_V \in R^{T\times d_V}$,然后利用注意力机制实现语言与视觉模态特征的跨模态交互,即

$$Q_L = X_L W_Q,\ K_V = X_V W_K,\ V_V = X_V W_V; \quad (1)$$

$$\mathrm{CroA}(Q_L, K_V, V_V) = \mathrm{softmax}\left(\frac{Q_L K_V^{\mathrm{T}}}{\sqrt{d_k}}\right) V_V. \quad (2)$$

其中: $W_Q \in R^{d_L\times d_k}, W_K \in R^{d_V\times d_k}, W_V \in R^{d_V\times d_v}$为语言与视觉特征映射权重; CroA表示语言与视觉模态特征的跨模态注意力实现算子.

其次,为强化目标模态语言的特征信息,通过残差连接的方式使语言模态信息的Q_L得到视觉模态信息的补充. 强化后的语言模态特征通过下式:

$$\tilde{X}_{V\to L} = \mathrm{LaNorm}\left(Q_L + \mathrm{CroA}\left(Q_L, K_V, V_V\right)\right) \quad (3)$$

进行标准化,得到$\tilde{X}_{V\to L} \in R^{T\times d_v}$.

$$X_{V\to L} = \mathrm{LaNorm}(\tilde{X}_{V\to L} + \mathrm{FFN}(\tilde{X}_{V\to L})), \quad (4)$$

$$\mathrm{FFN} = W_2(\mathrm{ReLU}(W_1 \tilde{X}_{V\to L} + b_1)) + b_2. \quad (5)$$

利用式(4)和(5)将不同模态交互后的模态信息进行前向传播,即将$\tilde{X}_{V\to L}$输入前馈网络并进行按层规范化,得到以视觉为辅助模态,以语言为目标模态的跨模态Transformer $V \to L$的输出$X_{V\to L} \in R^{T\times d_v}$. 该前馈网络中包括两个仿射变换,其中ReLU为激活函数.

同理可得到以视频为辅助模态、以语言为目标模态的跨模态Transformer $A \to L$的输出$X_{A\to L} \in R^{T\times d_A}$. 每对跨模态Transformer分别用于模拟不同模态之间的相互作用. 因此,对于不同目标模态,每对跨模态Transformer的融合结果由下式得到:

$$Y_L = \mathrm{Concat}(X_{V\to L}, X_{A\to L}), \quad (6)$$

$$Y_V = \mathrm{Concat}(X_{L\to V}, X_{A\to V}), \quad (7)$$

$$Y_A = \mathrm{Concat}(X_{L\to A}, X_{V\to A}). \quad (8)$$

从而得到$Y_L \in R^{T\times 2d_v}, Y_V \in R^{T\times 2d_v}, Y_A \in R^{T\times 2d_v}$,即为跨模态Transformer模块的输出.

2.2　多模态LSTM系统

虽然Transformer中的自注意力机制也可以作为一种捕获长距离依赖的手段,但最近研究[33]表明: Transformer编码器在深层结构中各位置输出是相似的,难以对时间维度信息建模. 跨模态Transformer模块虽然可完成多模态特征在空间上的初始交互,但由于在计算过程中缺乏过去时刻的特征信息,无法模拟时序上的语义关系. 为捕捉不同模态在时间序列上的向前与向后的依赖关系,结合双向LSTM较强的长距离语义捕获能力,设计了多模态BiLSTM系统,通过利用双向LSTM网络完成多模态序列的计算.

多模态BiLSTM系统以DEAN模型的前一模块(跨模态Transformer模块)的输出$Y_L \in R^{T\times 2d_v}, Y_V \in R^{T\times 2d_v}$及$Y_A \in R^{T\times 2d_v}$作为输入,在时间维度上完成更新,得到输出$T_L \in R^{T\times d_t}, T_V \in R^{T\times d_t}$以及$T_A \in R^{T\times d_t}$. 具体地,对于各输入模态$m \in \{L, A, V\}$(其中$L$表示语言模态,$A$表示音频模态,$V$表示视觉模态),当前时刻跨模态Transformer的输出特征Y_t^m,依次经过LSTM中的输入门、遗忘门和输出门与上一时刻记忆单元中的特征C_{t-1}^m进行比较,输出特征T_t^m. 在深度情感唤醒网络中,t时刻第m个模态的LSTM更新公式如下:

$$i_t^m = \sigma(W_i^m[h_{t-1}^m, Y_t^m] + b_i^m), \quad (9)$$

$$f_t^m = \sigma(W_f^m[h_{t-1}^m, Y_t^m] + b_f^m), \quad (10)$$

$$o_t^m = \sigma(W_o^m[h_{t-1}^m, Y_t^m] + b_o^m), \quad (11)$$

$$\tilde{c}_t^m = \tanh(W_{\tilde{c}}^m[h_{t-1}^m, Y_t^m] + b_{\tilde{c}}^m), \tag{12}$$

$$c_t^m = f_t^m \odot c_{t-1}^m + i_t^m \odot \tilde{c}_t^m, \tag{13}$$

$$T_t^m = o_t^m \odot \tanh(c_t^m). \tag{14}$$

其中: i_t^m, f_t^m, o_t^m 分别为 t 时刻第 m 个模态的输入门、遗忘门和输出门; $W_i^m, W_f^m, W_o^m, W_{\tilde{c}}^m$ 分别为变换中的参数矩阵; σ 为 Sigmoid 激活函数; $\odot$ 为 Hadamard 积.

2.3 多模态门控模块

本节构建了一个多模态门控模块(图3),用以模拟情感唤起模型的化学/神经激活结构. 多模态门控模块作为激活机制,不仅可以对带有时序信息的各目标模态进一步激活,还可以通过判断不同模态的重要性来控制每个目标模态的输出. 多模态门控模块的设计以自适应门控机制为基础,在多模态情感分析领域中自适应门控机制已经被证明是有效的. 如 RAVEN 模型[11],即通过自适应门控机制实现利用其他模态特征对其中一种模态进行增强的效果. DEAN 中的多模态门控模块首先通过自适应门控机制对不同输入模态的重要性进行权重分配,而后利用激活函数隐式实现多模态信息的融合.

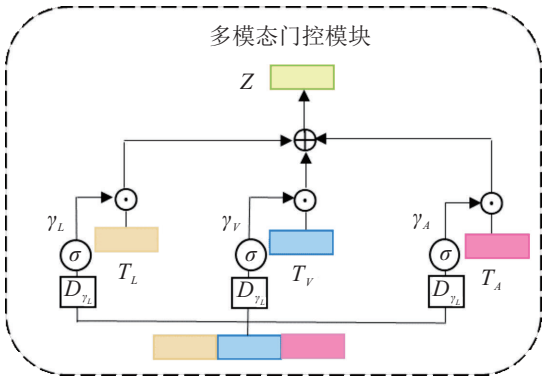


图 3 多模态门控模块

首先将 T_L、T_V、T_A 进行连接,分别输入3个前馈神经网络以学习得到各模态的权重向量; 然后通过下式控制各目标模态的输出:

$$T^{[L,V,A]} = \text{Concat}(T_L, T_V, T_A), \tag{15}$$

$$\gamma_L = D_{\gamma_L}(T^{[L,V,A]}), \tag{16}$$

$$\gamma_V = D_{\gamma_V}(T^{[L,V,A]}), \tag{17}$$

$$\gamma_A = D_{\gamma_A}(T^{[L,V,A]}); \tag{18}$$

最终完成各模态信息的融合,得到多模态门控模块的输出 $Z \in R^{T\times d_t}$,其中 $T^{[L,V,A]}$ 由 T_L、T_V、T_A 连接所得. 式(16)~(18)中 $\{\gamma_L, \gamma_V, \gamma_A\} \in R^{T\times d_t}$,表示目标模态门控以 $T^{[L,V,A]}$ 为输入,经 D_{γ_L}、D_{γ_V}、D_{γ_A} 三个独立的双层前馈神经网络学习得到.

3 实验及结果分析

3.1 数据集与实验环境

为验证本文所提出DEAN的性能,分别在多模态情感分析与多模态情绪识别两类任务上进行实验. 针对多模态情感分析任务,使用数据集CMU-MOSI和CMU-MOSEI. CMU-MOSI数据集收集了89位讲述者的视频,由2 199个短视频片段组成; CMU-MOSEI数据集选取了3 228个独白视频片段. CMU-MOSI数据集与CMU-MOSEI数据集在2分类任务中将每个片段标记为正面情绪或负面情绪. 7分类任务中标签为范围[−3,+3]的情感倾向得分,其中−3代表强烈负面情绪,+3代表强烈正面情绪. 多模态情绪识别的任务目标是通过多种模态的信息判断人的情绪状态. 针对此任务,采用IEMOCAP数据集收集了10名演员的302个会话视频, 视频中每个片段具有对应的情绪标注,包含愤怒、兴奋、恐惧、悲伤、惊讶、沮丧、高兴、失望和中性9种的情绪.

实验全部在服务器端完成计算,具体实验环境如下: CPU, 6×Xeon E5-2678 v3, GPU, RTX 2080 Ti; Python 3.7.6, CUDA 10.0, Pytorch 1.3.1, cudNN 7.6.

3.2 基准模型

选取目前流行的几种性能较高的代表性模型:

BC-LSTM[34] 是一个基于LSTM的多模态情感分析模型,可以在视频中捕获上下文信息;

TFN[9] 是一种张量融合模型,采用端到端的方式学习了特定模态以及跨模态两种动态的集成;

MFN[35] 利用特定注意力机制构造了一个多视图门控记忆模块,记忆单元随隐藏状态的变化而更新;

Graph-MFN[36] 在MFN模型上引入动态融合的新型可解释融合机制,增加了模型的可解释性;

RAVEN[11] 利用非语言特征动态地对语言特征建模,并改变词语的表示;

MulT[10] 利用Transformer结构对多模态序列交互过程建模.

3.3 评价指标与参数设置

为衡量DEAN的性能,采用分类正确率Acc(包括7分类正确率Acc7及2分类正确率Acc2),F1得分和平均绝对误差MAE三个指标来评估模型.

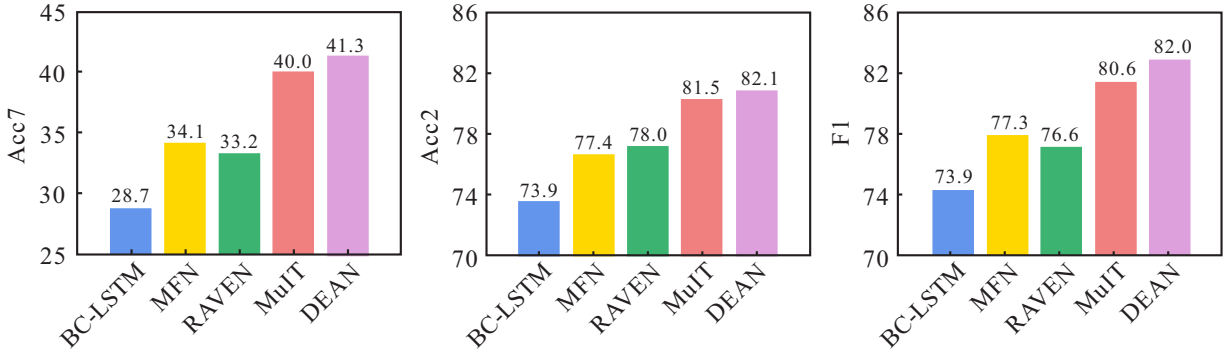
实验过程中数据集分别按70%、20%和10%的比例划分为训练集、测试集和验证集. 经多次交叉验证后,得到DEAN的参数设置如下:

批量大小 bs = 24, 学习率 lr = 0.001; 跨模态 Transformer 模块初始输入大小 $\text{size}_{X_L} = \text{size}_{X_V} = \text{size}_{X_A} = 30$; 各跨模态 Transformer 中 head 与 layer 数量 $\text{num}_{\text{head}} = 5, \text{num}_{\text{layer}} = 4$.

多模态 LSTM 系统中模态输入大小 $\text{size}_{Y_L} = \text{size}_{Y_V} = \text{size}_{Y_A} = 60$, 各模态 BiLSTM 层数 $\text{lay}_{Y_L} = \text{lay}_{Y_V} = \text{lay}_{Y_A} = 2$.

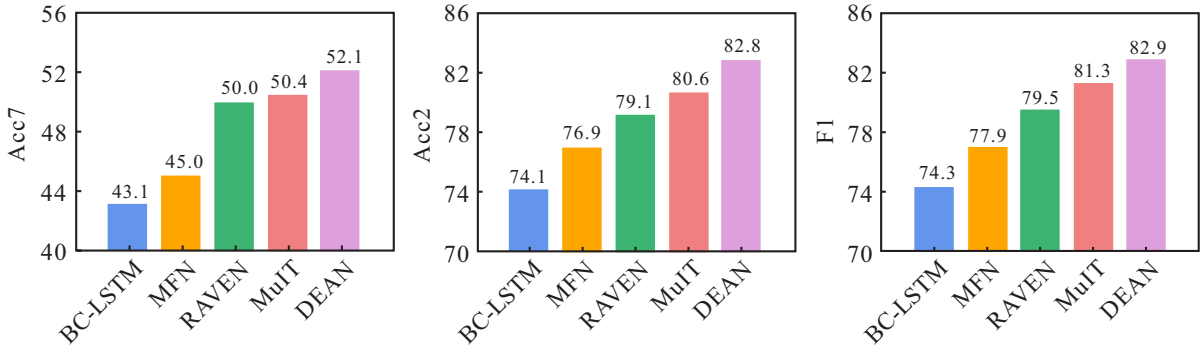
多模态门控模块输入 $\text{size}_{T_L} = \text{size}_{T_V} = \text{size}_{T_A} = 60$. 为了方便与基准模型进行比较, 模态对应门控尺寸的设置与模块输入尺寸保持一致, 即 $\text{size}_{\gamma_L} = \text{size}_{\gamma_V} = \text{size}_{\gamma_A} = 60$.

由图 4 和图 5 可以看出: 在相同数据集上, DEAN 在所有评价指标上都优于目前文献中最先进的方法. 图 5 表明在 CMU-MOESI 数据集上, Acc7 至少增加了 1.3 个百分点, MAE 也显著降低 (从 0.599 降到 0.573). 图 4 也给出了在 CMU-MOSI 数据集上类似的结果. 此外, 本文还在 IEMOCAP 数据集上进行了多模态情绪识别实验, 通过高兴、悲伤、愤怒、中性 4 种情绪的准确率, 以及 F1 得分来评估所提出的方法, 比较结果见表 1.



(a) 7 分类精度　　(b) 2 分类精度　　(c) F1分数

图 4　CMU-MOSI 数据集情感分类实验结果比较



(a) 7 分类精度　　(b) 2 分类精度　　(c) F1分数

图 5　CMU-MOESI 数据集情感分类实验结果比较

表 1　IEMOCAP 数据集上的实验结果比较

metric	happy		sad		angry		neutral	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BC-LSTM	83.1	81.7	82.1	81.7	85.0	84.2	66.1	64.1
MFN	90.2	85.8	**88.4**	86.1	87.5	86.7	72.1	68.1
RAVEN	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
MulT	**90.7**	88.6	86.7	86.0	87.4	87.0	**72.4**	70.7
DEAN (ours)	90.6	**89.2**	86.4	**86.3**	**88.1**	**87.7**	**72.4**	**71.5**

3.4　对比实验结果与分析

由表 1 可以看出, DEAN 模型除悲伤情绪类别的准确率低于 MFN 模型之外, 在高兴、愤怒、中性情绪类别上的表现均超越了当前最先进模型的结果.

3.5　消融实验研究

为分析 DEAN 中每个模块的影响、不同模态的重要性差异以及模态之间的相互作用, 本文在 CMU-MOSEI 数据集上进行了消融实验.

3.5.1　各模块对 DEAN 模型的影响

DEAN 完整模型由 3 个模块组成, 即多模态双向 LSTM 系统, 跨模态 Transformer 模块和多模态门控模块. 本文以 DEAN 完整模型为基准模型进行如下的消融实验.

DEAN w/o GATE: 在完整模型的基础上删除输出前的多模态门控模块;

DEAN w/o LSTMs: 在完整模型的基础上删除多

模态双向LSTM系统;

DEAN w/o LSTMs & GATE: 在完整模型的基础上只保留跨模态Transformer模块.

表2给出了CMU-MOSEI数据集上模块组合的消融实验结果. 实验结果显示完整的DEAN模型的实验效果最好,缺少不同模块会对实验结果造成不同影响. 以Acc7指标分析,相比于完整DEAN模型52.1 %的准确率,缺少多模态门控模块准确率下降了1.7个百分点,而缺少多模态BiLSTM模块,模型准确率下降了1.3个百分点. 由此可知,在情感分析任务中过去时刻情感信息对当前时刻的情感存在着重要影响.

表 2 CMU-MOSEI数据集上模块组合的消融实验结果

	Acc7	Acc2	F1	MAE
DEAN	52.1	82.8	82.9	0.573
DEAN w/o GATE	51.4	82.0	82.3	0.591
DEAN w/o LSTMs	50.8	82.2	81.6	0.596
DEAN w/o LSTMs & GATE	50.3	80.9	80.9	0.599

3.5.2 不同模态的重要性

为进一步探讨不同模态在多模态任务中的重要性,分别以语言、音频和视觉模态作为独立输入进行实验. 将各模态特征分别输入对应的语言模型、音频和视觉模型中比较. 模态消融研究结果如表3所示. 由表3的结果观察到: 与视觉和音频相比,单独使用语言模态会取得更好的实验效果. 这可能是因为语言被认为是情感分析的视觉和听觉模式的支点. 同时,这一观察也验证了多模态门控模块的重要性,即各模态对于识别的最终结果影响是不相同的. 最后,输入3种模态特征的完整模型明显优于仅输入单一模态的模型.

表 3 CMU-MOSEI数据集模块消融实验结果

metric	Acc7	Acc2	F1	MAE
language	47.2	77.4	78.5	0.652
audio	44.7	65.9	70.0	0.757
vision	43.0	65.2	69.8	0.777

3.5.3 模态间的相互作用

为研究模态之间的交互作用,以双模态和三模态的形式进行消融研究实验,分别以语言、视觉和听觉为目标模态,观察辅助模态和目标模态之间的交互作用. 结果如表4所示,其中箭头指向的方向表示目标模态.

由表4的结果可以得出结论: 辅助模态可以提高目标模态的性能. 两种辅助模态的实验效果好于单一辅助模态,其中各模态相互作用完整组合后(即DEAN模型)效果最好. 此外,以语言为目标模态的表现都优于以视觉或听觉为目标模态的表现. 表4的结果表明,以语言为目标,其他两种模态作为辅助模态的效果要好于以音频或视觉为目标模态模型的效果. 语言作为辅助模态也起着至关重要的作用. 在没有语言模态特征输入下,与$L \to V$相比,$A \to V$的Acc7降低了4.9 %,验证了语言模态信息在多模态情感分析中的重要性.

表 4 CMU-MOSEI模态相互作用的消融实验结果

metric	Acc7	Acc2	F1	MAE
$V \to L$	49.7	78.5	79.6	0.593
$A \to L$	50.0	78.1	79.6	0.596
$L \to V$	49.2	77.5	79.4	0.626
$A \to V$	44.3	69.6	71.7	0.755
$L \to A$	48.1	77.5	79.1	0.631
$V \to A$	44.5	69.2	71.9	0.752
$V, A \to L$	51.4	80.6	81.4	0.579
$L, A \to V$	51.1	79.8	80.8	0.597
$L, V \to A$	50.8	78.4	80.4	0.580
DEAN	52.1	82.8	82.9	0.573

4 结 论

受认知科学中情感唤醒模型的启发,本文提出了一种处理多模态情感分析与情感识别任务的完整框架: 深度情感唤醒网络(DEAN). DEAN模型模拟人类处理多通道信息的过程,通过在跨模态Transformer基础上引入时序信息,使其能够实现对情感连续性的建模,并通过多模态门控模块实现了针对不同模态的输出信息的控制. 在3个经典数据集上进行的多模态情感分析和情绪识别的综合对比实验与消融实验的结果表明,DEAN模型在所有数据集上几乎都达到或超越当前先进的情感识别模型的性能.

多模态情感分析领域任务与脑神经科学中多感觉整合[37]研究具有很高的相关性,未来将结合认知计算的先进方法[38-39]进行多模态情感分析任务的相关研究.

参考文献(References)

[1] LaBar K S, Cabeza R. Cognitive neuroscience of emotional memory[J]. Nature Reviews Neuroscience, 2006, 7(1): 54-64.

[2] Lehrer S F, Xie T. The bigger picture: Combining

econometrics with analytics improves forecasts of movie success[J]. Management Science, 2022, 68(1): 189-210.

[3] O'Connor B, Balasubramanyan R, Routledge B, et al. From tweets to polls: Linking text sentiment to public opinion time series[C]. Proceedings of the International AAAI Conference on Web and Social Media. Atlanta, 2014: 122-129.

[4] da Cao, Miao L H, Rong H G, et al. Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities[J]. Knowledge-Based Systems, 2020, 203: 106114.

[5] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, 2017: 873-883.

[6] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J/OL]. 2018, arXiv: 1806.00064.

[7] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2015: 1449-1457.

[8] Baltruaitis T, Banda N, Robinson P. Dimensional affect recognition using continuous conditional random fields[C]. The 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Shanghai, 2013: 1-8.

[9] Zadeh A, Chen M H, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J/OL]. 2017, arXiv: 1707.07250.

[10] Tsai Y H H, Bai S J, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]. Proceedings of the Conference on Association for Computational Linguistics. Florence, 2019: 6558-6569.

[11] Wang Y S, Shen Y, Liu Z, et al. Words can shift: Dynamically adjusting word representations using nonverbal behaviors[J]. Proceedings of the AAAI Conference on Artificial Intelligence AAAI Conference on Artificial Intelligence, 2019, 33(1): 7216-7223.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems. Boston: MIT Press, 2017: 5998-6008.

[13] Delbrouck J B, Tits N, Brousmiche M, et al. A transformer-based joint-encoding for emotion recognition and sentiment analysis[C]. The 2nd Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Seattle: Association for Computational Linguistics, 2020: 1-7.

[14] Zadeh A, Zellers R, Pincus E, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J/OL]. 2016, arXiv: 1606.06259.

[15] Gkoumas D, Li Q C, Lioma C, et al. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis[J]. Information Fusion, 2021, 66: 184-197.

[16] Lindsay P H, Norman D A. Human information processing: An introduction to psychology[M]. London: Academic Press, 2013: 191-253.

[17] Pons G, Masip D. Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition[J]. IEEE Transactions on Cybernetics, 2020, 99: 1-8.

[18] Acar E, Hopfgartner F, Albayrak S. A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material[J]. Multimedia Tools and Applications, 2017, 76(9): 11809-11837.

[19] Zhong S H, Wu J X, Jiang J M. Video summarization via spatio-temporal deep architecture[J]. Neurocomputing, 2019, 332: 224-235.

[20] Zhu Y Y, Tong M, Jiang Z B, et al. Hybrid feature-based analysis of video's affective content using protagonist detection[J]. Expert Systems with Applications, 2019, 128: 316-326.

[21] Zhu X G, Li L, Zhang W G, et al. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition[C]. Proceedings of the 26th International Joint Conference on Artificial Intelligence. San Francisco, 2017: 3595-3601.

[22] Sivaprasad S, Joshi T, Agrawal R, et al. Multimodal continuous prediction of emotions in movies using long short-term memory networks[C]. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. Yokohama, 2018: 413-419.

[23] Gui D D, Zhong S H, Ming Z. Implicit affective video tagging using pupillary response[C]. International Conference on Multimedia Modeling. Cham: Springer, 2018: 165-176.

[24] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion[C]. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Berlin: Springer, 2014: 153-162.

[25] Wu Z Y, Cai L H, Meng H. Multi-level fusion of audio and visual features for speaker identification[C] International Conference on Biometrics. Berlin: Springer, 2006: 493-499.

[26] Vielzeuf V, Pateux S, Jurie F. Temporal multimodal

fusion for video emotion classification in the wild[C]. Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasow, 2017: 569-576.

[27] Gönen M, Alpaydn E. Multiple kernel learning algorithms[J]. The Journal of Machine Learning Research, 2011, 12: 2211-2268.

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

[29] Mai S J, Xing S L, Hu H F. Locally confined modality fusion network with a global perspective for multimodal human affective computing[J]. IEEE Transactions on Multimedia, 2020, 22(1): 122-137.

[30] Pham H, Liang P P, Manzini T, et al. Found in translation: Learning robust joint representations by cyclic translations between modalities[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, 2019: 6892-6899.

[31] Pham H, Manzini T, Liang P P, et al. Seq2Seq2Sentiment: multimodal sequence to sequence models for sentiment analysis[C]. Proceedings of Grand Challenge and Workshop on Human Multimodal Language. Melbourne, 2018: 53-63.

[32] 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. New Orleans, 2018: 5642-5649.

[33] Li X T, Li G L, Liu L M, et al. On the word alignment from neural machine translation[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 1293-1303.

[34] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[35] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[J/OL]. 2018, arXiv: 1802.00927.

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

[37] Stein B E, Stanford T R. Multisensory integration: Current issues from the perspective of the single neuron[J]. Nature Reviews Neuroscience, 2008, 9(4): 255-266.

[38] Wu E Q, Xiong P W, Tang Z R, et al. Detecting dynamic behavior of brain fatigue through 3-D-CNN-LSTM[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(1): 90-100.

[39] Wu E Q, Hu D W, Deng P Y, et al. Nonparametric Bayesian prior inducing deep network for automatic detection of cognitive status[J]. IEEE Transactions on Cybernetics, 2021, 51(11): 5483-5496.

作者简介

张峰 (1976−), 女, 副教授, 博士, 从事机器学习、智能决策等研究, E-mail: fengzhang@hbu.edu.cn;

李希城 (1996−), 男, 硕士生, 从事机器学习的研究, E-mail: 763943848@qq.com;

董春茹 (1980−), 男, 副教授, 博士, 从事深度学习、图像处理等研究, E-mail: dongcr@hbu.edu.cn;

花强 (1973−), 男, 教授, 从事机器学习、智能计算等研究, E-mail: huaq@hbu.edu.cn.

(责任编辑：孙艺红)