Journal of Computer Applications
计算机应用
ISSN 1001-9081
CODEN JYIIDU
2020-12
http://www.joca

文章编号:1001-9081(****)**-0000-00
doi:10.11772/j.issn.1001-9081.**2020091416**

基于多任务学习的时序多模态情感分析模型

章荪，尹春勇*

(南京信息工程大学 计算机与软件学院，南京 210044)

(*通信作者电子邮箱 yinchunyong@hotmail.com)

摘　要: 针对时序多模态情感分析中存在的单模态特征表示和跨模态特征融合问题，结合多头注意力机制，提出一种基于多任务学习的情感分析模型。首先，结合卷积神经网络、双向门控循环神经网络和多头自注意力实现了对时序单模态的特征表示；然后，利用多头注意力实现跨模态的双向信息融合；最后，基于多任务学习思想，添加额外的情感极性分类和情感强度回归任务作为辅助，提升情感评分回归主任务的综合性能。通过实验结果可以证明，所提模型在二分类准确度指标上，相较于多模态分解模型，在 CMU-MOSEI 和 CMU-MOSI 多模态数据集上分别提高了 7.8 和 3.1 个百分点。因此，该模型适用于多模态场景下的情感分析问题，能够为商品推荐、股市预测、舆情监控等应用提供重要的决策支持。

关键词: 情感分析；多模态；多任务学习；序列学习；特征融合

中图分类号:TP391.1　　**文献标志码：A**

Sequential multimodal sentiment analysis model based on multi-task learning

ZHANG Sun, YIN Chunyong*

(1. School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing Jiangsu 210044, China)

Abstract: Considering the issues of unimodal feature representation and cross-modal feature fusion in sequential multimodal sentiment analysis, a multi-task model was proposed based on multi-head attention mechanism. Firstly, convolution neural network, bidirectional gated recurrent neural network and multi-head self-attention were integrated for sequential unimodal feature representation. Secondly, bidirectional cross-modal information was fused by multi-head attention. Finally, sentiment polarity classification and intensity regression were designed as additional auxiliary tasks to improve the performances of main sentiment score regression task based on multi-task learning. Experimental results can demonstrate that proposed model has improved accuracy rate by 7.8 and 3.1 per cents respectively on CMU-MOSEI and CMU-MOSI datasets, compared with multimodal factorized model. Therefore, the proposed model has provided valuable decision supports for product recommendation, stock market forecasting, public opinion monitoring and other relevant applications.

Keywords: sentiment analysis; multimodal; multi-task learning; sequential learning; feature fusion

0　引言

情感分析主要涉及检测、分析和评估用户面对不同事件、问题、服务时所产生的心理状态，它是实现智能化人机交互的必要条件[1]。社交网络作为新兴的信息媒体，允许用户上传和分享日常的生活经历和观点看法。这些自用户端发布的数据含有丰富的情感信息，能够为情感分析提供重要的数据支持[2]。因此情感分析相关的工作大多以社交网络为背景，利用用户生成数据训练情感分析模型。

现有的情感分析研究主要集中于单一的文本模态，它伴随着统计学习和人工智能技术的发展得到了不断地完善。文本情感分析的关键在于构建有有效的文本特征表示。早期方法通常基于词汇的情感信息，提取词语统计特征作为文本表示，利用机器学习方法实现进一步的分类和预测。而自深度学习兴起后，研究者提出利用卷积神经网络(Convolution Neural Network, CNN)或循环神经网络(Recurrent Neural Network, RNN)端到端提取文本的空间和时序信息，或是利用预训练的语言模型将高维的稀疏特征映射到低维的语义空间，学习文本的嵌入表示。深度学习解决了传统统计方法中存在的维度爆炸和特征稀疏问题，但是这些方法通常只关注于单一的模态信息，不能适应多模态的社交网络环境。

每一种信息的来源或形式都可以看作是一种模态，社交网络正是由文本、图像、语音等多种模态构成的复杂环境[3]。

收稿日期: 2020-09-14;　修回日期:2020-10-24;　录用日期:2020-10-30。
基金项目：国家自然科学基金（61772282）。
作者简介:章荪(1994-)，男，安徽六安人，博士研究生，主要研究方向：深度学习、情感分析、文本分类；尹春勇(1977-)，男，山东潍坊人，教授，博士，博士生导师，主要研究方向：网络空间安全、大数据挖掘及隐私保护、人工智能及新型计算。

例如在博客和商品评论场景中，用户上传的信息通常包括文字和图像两部分内容，两种模态之间具有一定的语义和情感相关性。图像内容的信息能够辅助增强文本内容的情感表达，有效缓解可能出现的词语歧义、语义模糊等问题。Truong 等[4]关注于图文模态之间的特征融合问题，并指出多模态情感分析能够利用不同模态信息的一致性和互补性实现精准的情感预测。而 Verma 等[5]则进一步指出模态内部自身的独有特征也不能被忽略。因此，多模态场景下的情感分析工作需要解决模态的异质性和异构性问题，挖掘模态内部自身独有的特征信息以及模态之间的交互信息。

在以油管、抖音为代表的视频流媒体中，用户上传的视频可以看作是文字、图像、语音三种模态信息混合的时序数据。不同于静态的图文混合场景，模态之间的交互发生在时间尺度上，并且模态内部具有时序特征[6]。因此，时序多模态情感分析需要解决两点问题：单模态的时序特征表示问题和跨模态的时序特征融合问题。Pham 等[7]基于机器翻译的序列到序列模型(Sequence to Sequence, Seq2Seq)，利用循环神经网络提取各模态的时序特征，再利用编码-解码过程学习模态之间的关联性，以编码训练后的上下文特征作为跨模态的融合特征表示。Mai 等[8]提出的模态转换方法同样基于机器翻译模型，借助对抗训练提供编码器的推断能力，学习更好的单模态特征表示，再利用图融合网络分级融合不同模态的信息。此类基于机器翻译和编码-解码结构的方法，能够解决模态缺失和噪声干扰的问题，但是在各模态信息较为完整时，情感分类准确度通常略有较低。

Tsai 等[9]利用多头注意力机制计算两两模态组合之间的关联程度，提出多模态 Transformer(Multimodal Transformer, MulT)模型，能够直接处理未对齐的模态序列，但是该方法未充分挖掘模态自身的时序信息，并且在预测时仅使用融合后的特征，忽略了模态内部所独有的特征。因此，为了提取单模态内部的时序信息，本文提出集成卷积网络、双向门控网络(Bidirectional Gated Recurrent Unit, BiGRU)和多头自注意力(Multi-Head Self-Attention)的时序特征表示方法。Kim 等[10]最早提出了基于卷积神经网络的文本时序特征提取方法，TextCNN(Text Convolution Neural Network)模型能够实现类似于 N-Gram 模型的效果，利用多个一维卷积核提取短语级的特征信息。在时序特征提取过程中，本文还利用卷积网络实现了模态特征维度的统一，方便后续特征融合阶段的注意力计算。双向循环神经网络能够发现序列数据前向和后向的关联性，而多头自注意力利用注意力机制提取上下文信息，二者都被广泛地应用于序列建模问题中，用于提取时序特征。二者的区别在于前者公平地对待每一个序列位置上的数据，而后者则为每个位置上的数据分配不同的注意力权重。循环神经网络因为隐藏神经元的遗忘门机制和维度的有限，无法储存长期的记忆信息，不适用于过长的序列数据。而注意力机制与所有的序列输入建立连接，能够获得全局的上下文信息。将注意力机制引入循环神经网络能够解决其存在的局限性，更好地提取序列数据的时序特征。此外，为了挖掘模态之间的交互关系，本文基于多头注意力机制提出了跨模态时序特征融合方法，发现模态组合之间双向的对应关系，实现了跨模态信息的融合。

在获得单模态特征表示及跨模态的融合特征后，本文基于多任务学习(Multi-Task Learning, MTL)设计下游任务框架。以情感评分回归作为主任务，额外添加情感极性分类和情感强度回归作为辅助任务，帮助上游模型提取更具区分度和泛化性的特征。Tian 等[11]最早将多任务学习机制应用于多模态情感分析问题，依据情感评分回归主任务，设计情感极性和强度分类作为辅助任务。Akhtar 等[12]同样基于多任务学习的思想，提出 CIM-MTL(Contextual Inter Modal-Multi Task Learning)模型，设计情感极性二分类任务辅助实现细粒度的情感分类主任务。考虑到每种模态信息在不同任务中具有不同的贡献度和重要性，本文为下游模型添加任务专属的独立评分模块，按照任务需要计算每个共享特征的重要性，构建任务专属的融合特征表示。本文主要贡献有三点：(1) 提出单模态时序特征表示方法，通过集成卷积网络、双向门控神经网络和多头自注意力机制，充分挖掘序列数据的内部时序信息；(2) 提出跨模态特征融合方法，基于多头注意力机制，在时间尺度上挖掘模态之间的交互关系，融合双向注意力加权结果；(3) 提出任务专属特征融合方法，为下游多任务学习模型添加独立的评分模块，根据具体任务目标，为共享特征表示自适应分配权重系数，构建任务专属的融合特征。

全文的组织如下所示：第一节将简要回顾多模态情感分析和多任务学习相关的背景知识；第二节中将详细介绍本文提出的改进模型；第三节中利用两个公开的多模态数据集对本文模型进行分析和检验，通过定性和定量的实验证明了本文模型能够实现更好的情感分类效果；第四节对全文进行总结和分析，讨论本文方法的优劣并介绍未来的研究计划。

1 相关工作

1.1 情感分析

“情感”一词不仅指代人类具体的一种情感状态，更是泛指一切感官、机体、心理以及精神的感受，能够借由语言进行传递和表达。分析和理解用户的情感状态是实现人工智能、情感计算和人机交互的必要条件。在不同的情感分析问题中，研究者通常使用“sentiment”或“emotion”这两个术语来表示情感，前者通常与情感极性分类或回归任务相关，将情感粗略的划分为积极和消极两种状态（部分研究中会添加中性状态），分析用户主观感觉的倾向性，或是以实数情感评分度量用户的情感状态。而后者则一般涉及到细粒度具体的情感类别分类，通常基于心理学和认知学的情感表示模型，将情感状态归纳到不同的类别[13]。常用的情感表示模型如表 1 所示。

在 Hovy 等[14]的研究工作中，情感分析被定义为判断说话者或写作者对某个特定主题或文档全部内容的态度，而这种态度包含人的主观判断、情感状态或某种情感交流，他们认为情感分析包含了观点挖掘、情感分类、极性分类等一系列问题，“sentiment”和“emotion”可以统一为主体对特定主题产生的主观感觉。而 Munezero 等[15]则认为“sentiment”比“emotion”更加稳定且具有更强的倾向性，是针对特定对象产生的。实际上，在具体的应用中二者的边界是很模糊的，本文根据任务目标的不同对二者进行区分。本文在情感极性二分类任务中，使用“sentiment”表示粗粒度的情感倾向，将情感极性粗略的划分为积极和消极两种状态。在七分类任务中，则使用“emotion”表示具体的细粒度情感，采用七级李克特量表作为情感表示模型。

表 1 情感表示模型

Tab. 1 Representative emotion models

模型	情感状态
埃克曼六种基本情感	愤怒、厌恶、恐惧、快乐、悲伤、惊喜
米克尔八种情感	厌恶、悲伤、恐惧、愤怒、娱乐、满足、敬畏、兴奋
普鲁契克情感色轮	快乐、信任、恐惧、惊讶、悲伤、期待、愤怒、厌恶（三个等级）
七级李克特量表	非常否定、否定、比较否定、一般、比较肯定、肯定、非常肯定

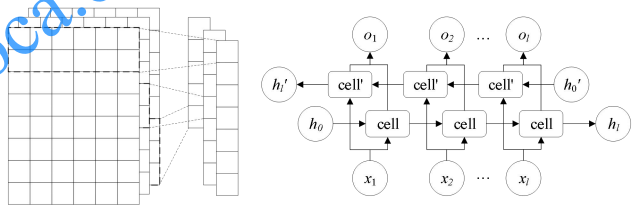
1.2 多模态情感分析

早期的情感分析主要面向单一的文本数据，利用自然语言处理、统计分析、计算语言学等技术，对携带情感信息的文本内容进行处理、分析、归纳和推理。

文本情感分析方法得益于文本分类技术的发展得到了不断地完善和改进。在面向单模态的情感分析研究中，文本内容通常被认为能够更好地表达情感和态度，因为词语本身包含了大量情感相关的信息，而图像和语音在情感表达上存在着语义混淆的可能。文本分类与文本情感分析都需要提取文本的语义信息，因此二者在技术上具有一定的相似性，而图像情感分析与图像分类有着本质的不同，图像分类模型中提取的纹理视觉特征不能表示图像的情感信息，它需要更高等级的抽象来发现潜在的语义信息。Borth 等[16]首次提出利用形容词-名词对组合作为中级特征表达图像的语义信息，再利用分类器预测情感类别与词语组合之间的关联性。Guillaumin 等[17]发现结合与图像对应的文本内容能够帮助理解图像传达的语义信息，实现更精准的图像分类效果，这启发了更多的研究者尝试引入更多的模态信息，也使得多模态学习得到了持续的关注。

多模态学习能够将声觉和视觉内容与相应的文本信息进行关联，使得非文本信息能够被更好地理解。而非文本信息也能够从不同的视角赋予文本更多的含义，强化文本的情感表达。与传统静态的图文情感分析不同，视频数据可以分解成文本、语音、图像三种模态信息，每种模态都是一个时间序列，这种由多个时间序列混合而成的数据可以称为时序多模态。人类的语言同样是一个多模态的时序过程，在面对面交谈时，声音变化、面部动作和谈话内容都是时变的，这些信息都能够传递说话者的情感和态度。时序多模态情感分析存在着表示、转换、对齐和融合问题[18]，但一般而言，后三种问题可以总结为对跨模态交互关系的挖掘。因此，时序多模态情感分析面临的主要问题有两点，即单模态的时序特征表示和跨模态的时序特征融合问题。

首先，时序模态的特征表示方法需要发现模态内部不同时刻数据之间的关联性。常见的方法通常是利用卷积神经网络[19]或双向循环神经网络[20-21]提取时序特征。TextCNN 模型中提出了使用高度不同，宽度与序列数据维度相同的一组卷积核来提取序列的局部特征，如图 1(a)所示。这些高度不同的卷积核能够发现相邻时刻数据的关联性，用于文本序列时可以实现类似于 N-Gram 模型的效果，提取到短语级的特征。循环神经网络则通过模拟大脑的记忆、遗忘和更新，按照输入数据的顺序提取隐藏特征，作为序列新的特征表示。如图 1(b)所示，双向循环神经网络则可以进一步发现前向和后向的序列特征，被广泛应用于序列建模问题。



(a) 卷积神经网络　　(b) 双向循环神经网络

图 1 序列特征提取方法

Fig. 1 Sequential feature extraction methods

跨模态的特征融合需要整合多种模态的信息，发现模态之间存在的交互关系，而时序模态的交互是发生在时间尺度上的，即模态之间在每个时刻上都存在着一定的关联。常用的模态融合方法按照融合的阶段不同可以划分为两种：早期表示融合与晚期决策融合。决策融合通常是在获得每种模态的特征表示后，利用每种模态信息进行独立地预测，再经由加权、多数投票等处理获得最终的决策结果[22]。这种方法与集成学习相似，能够充分利用每种模态所独有的特征，具有较好的泛化性，但忽略了模态之间的关联性。

早期表示融合则是当前多模态学习关注的重点，一种常用的方法[23]则是直接拼接每种模态的特征，构建统一的联合表示进行预测。这种方法简单有效，但是只能获得浅层的信息，无法深度挖掘模态的交互关系，提取更抽象的特征。Zadeh 等[24]在 2017 年首次提出张量融合方法，利用向量的笛卡尔内积作为融合特征表示，这种方法能够同时捕获模态内部和模态之间的交互关系，提取单模态、双模态和三模态的特征，

但是具有较高的计算复杂度，随后 Liu 等[25]基于矩阵的低秩分解提出了高效的张量融合方法，Liang 等[26]则将该方法推广到时序多模态融合问题上。在注意力机制被提出后，基于注意力加权的表示融合方法得到了快速的发展。CIM-MTL 模型利用点乘注意力计算两个模态之间的相似性，再利用门控机制混合原始信息与融合信息。Yu 等[27]提出了基于多头注意力机制的单向特征融合方法，利用多个注意力头在不同子空间和位置上发现模态之间的关联性，而 MulT 模型则是多头注意力在时序多模态数据上的拓展。门控机制[28]可以看作是一种特殊的注意力机制，二者同样是利用神经网络学习权重系数，再经由加权求和获得融合的特征表示。

1.3 多任务学习

广义而言，在学习过程中同时优化多个损失函数都可以被认为是多任务学习，它的形式多样，联合学习、自主学习、辅助任务学习都可以被纳入其中。Caruana 等[29]最先定义了多任务学习的目标，即多任务学习利用包含在相关任务训练信号中的特定领域的信息来改进泛化能力。多任务学习具有一定的理论和实际意义，从人类学习的角度而言，人类通常会利用从相关任务学习到的知识帮助学习新的技能。以教育学的角度来看，人类通过学习相关任务来获得必要的技能，以支持掌握更加复杂的技术。

在深度神经网络中，多任务学习的实现通常采用两种参数共享机制：硬参数共享和软参数共享，如图 2 所示。前者是多任务学习中最常用的方法，它在所有的任务之间共享全部的隐藏层及其参数，仅保留最后几个任务专属的特定层。这种方法能够有效地降低过拟合的风险，因为在硬参数共享方法中，模型学习到适合所有任务的特征表示是困难的，这也能够降低在原始任务上过拟合的风险。后者则为每项任务都设置完整的模型和参数，但是会对任务模型的参数添加正则化约束，提高参数之间的相似性。

多任务学习在某种程度上实现了数据增强的效果，因为所有的任务都含有一定的噪声，在单个任务上训练模型时，期望的目标是能够学习到与该任务相关，并去除噪声干扰的特征表示。由于不同的任务具有不同的噪声模式，所以当一个模型同时学习多个任务时，就能够获得忽略多种噪声模式，学习到更具泛化性的特征表示。当一个任务含有大量的噪声或数据量有限并且维度过高，模型将难以提取到有效的信息，学习到相关的特征表示。而多任务学习则可以帮助模型将注意力集中在重要的特征上，因为其它相关的任务能够为这些特征的重要性提供额外的证据。此外，不同的特征在不同的任务上的学习难易程度不同。一些重要的特征可能在特定的任务上更容易被模型学习，而在其他的任务上可能由于复杂的交互方式或其他特征的干扰阻碍了模型的学习，多任务学习则可以利用多任务训练的优势提高模型的学习能力。

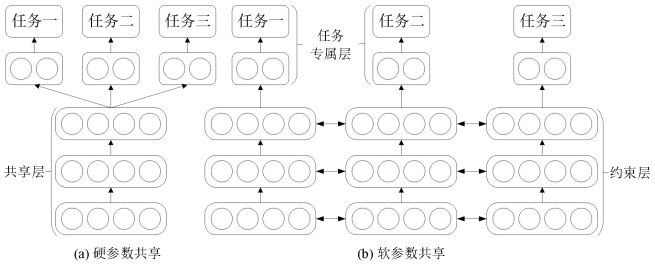


图 2 多任务学习的参数共享机制

Fig. 2 Parameter sharing mechanisms in MTL

现有的多模态情感分析工作中，集成多任务学习的方法大多采用硬参数共享方法，即在主任务和辅助任务之间共享上游的多模态融合网络和特征表示，并为每项任务设置专属的输出层及激活函数。本文同样基于硬参数共享机制，利用多任务学习的优点，学习更具泛化性的共享特征，具体的方法介绍见第二节。

2 基于多任务学习的时序多模态情感分析

本文的研究目标是实现对时序多模态数据的情感分析，所有的工作都是在 Zadeh 等[30-31]提出的 CMU-MOSI(CMU Multimodal Opinion level Sentiment Intensity) 和 CMU-MOSEI(CMU Multimodal Opinion Sentiment and Emotion Intensity)数据集上开展的。数据集中的每个样本 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_L\}$ 都是一个长度为 L 的时间序列，它可以分解为文本(T)、语音(A)、图像(V)三种序列模态 $\boldsymbol{X}=(\boldsymbol{X}^T,\boldsymbol{X}^A,\boldsymbol{X}^V)$。每个样本对应一个表示情感状态的实数评分 $y\in[-3,3]$，情感分析的目标是利用已有的数据样本训练一个模型，正确预测未知样本对应的评分。本文提出的多模态情感分析模型分为上游特征表示和下游多任务学习两部分，其中上游特征表示模型结构如图 3 所示，包含单模态时序特征表示和跨模态时序特征融合。

2.1 时序单模态特征表示

首先为了挖掘模态内部所独有的特征，并提取序列模态的时序信息，本文提出集成卷积神经网络、双向门控循环神经网络、多头自注意力机制的单模态时序特征表示方法。

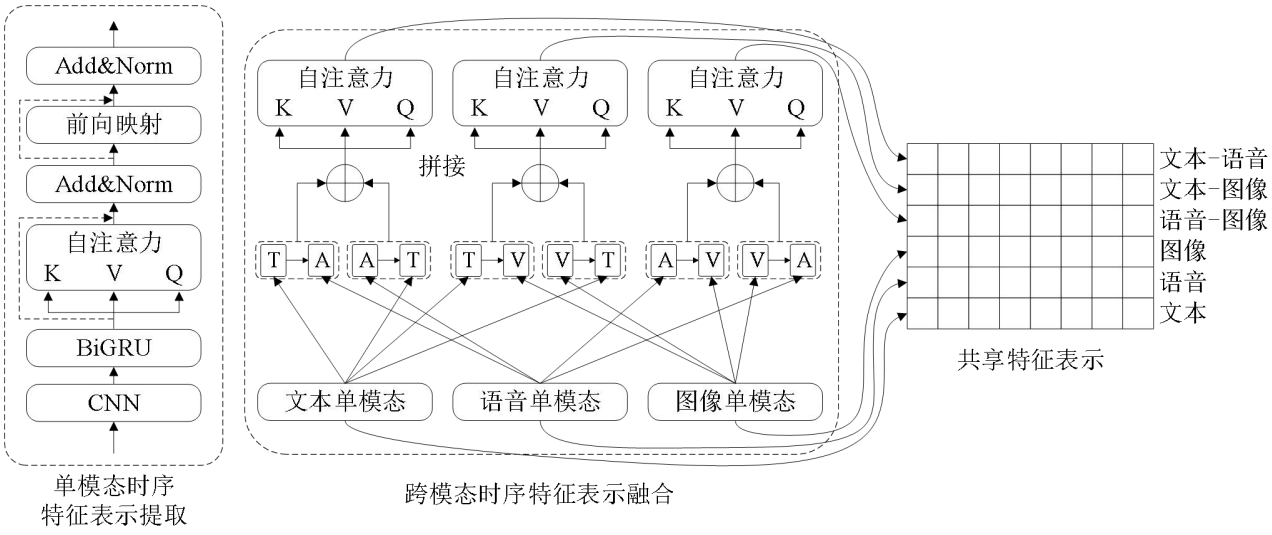


图 3 上游特征表示模型结构

Fig. 3 Architecture of upstream feature representation model

CNN 被证明能够提取序列的局部信息，具有滑动窗口和 N-Gram 模型相似的功能。在 MulT 模型中，CNN 还被用于统一各种模态的维度。本文使用一组固定高度，宽度与序列维度 $d_k;k\in\{T,A,V\}$ 相同的卷积核提取局部信息。如图 4 所示，经过 CNN 处理后的数据依然是一个时间序列，但维度被统一为卷积核的数量 d=#channel。在设置卷积核步长为 1 且不使用填充时，原始的时间序列长度会被缩短，这也有助于加速后续循环神经网络的训练，缩小注意力矩阵的形状。

$$\boldsymbol{X}_{conv}^{k}=\mathrm{Conv}\left(\boldsymbol{X}^{k}\right)\in\mathbf{R}^{d\times(L-H+1)};k\in\{T,A,V\} \tag{1}$$

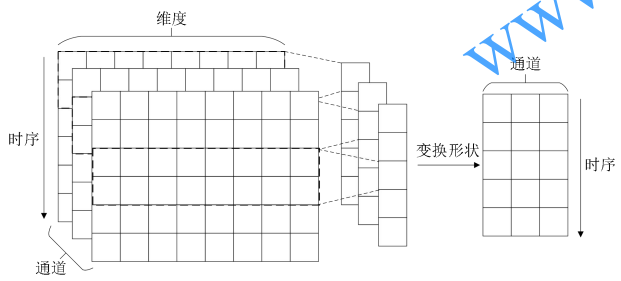


图 4 卷积神经网络提取局部时序信息

Fig. 4 Local temporal information extracted by CNN

CNN 处理后的数据将继续输入到 BiGRU 中，通过不断地更新隐藏状态，提取时间序列的高阶时间特征。设置隐藏状态的维度为 d_h，提取每个时刻对应的双向隐藏状态作为新的特征，因此 BiGRU 处理后的数据形状为 $2d_h\times(L\text{-}H\text{+}1)$。

$$\boldsymbol{H}^{k}=\mathrm{BiGRU}\left(\boldsymbol{X}_{conv}^{k}\right);k\in\{T,A,V\} \tag{2}$$

多头自注意力机制利用多个注意力头在不同的子空间内计算查询和索引向量之间的相似度，提取更加丰富的上下文信息。每个注意力头的计算公式如下所示：

$$\begin{aligned}Head_i\left(\boldsymbol{H}^{k}\right)=\mathrm{softmax}((\boldsymbol{W}^{Q}\boldsymbol{H}^{k})^{\mathrm{T}}(\boldsymbol{W}^{K}\boldsymbol{H}^{k})\\ /\sqrt{2d_h/M})(\boldsymbol{W}^{V}\boldsymbol{H}^{k})^{\mathrm{T}}\end{aligned} \tag{3}$$

其中，M 为注意力头的数量，$\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$ 和 $\boldsymbol{W}^{V}$ 分别是对应的映射矩阵，将原始数据映射到不同的低维空间。拼接所有注意力头的输出获得完整的输出结果：

$$MATT\left(\boldsymbol{H}^{k}\right)=\mathrm{concat}\left(Head_1,\cdots,Head_M\right) \tag{4}$$

多头自注意力输出的数据与查询矩阵逐元素累加，利用层归一化处理(Layer Normalization, LN)，避免数值过大而引起梯度爆炸问题 $\boldsymbol{SA}^{k}$=LN($\boldsymbol{H}^{k}$+$MATT$($\boldsymbol{H}^{k}$))。在经过全连接网络(Fully Connected Network, FC)映射和逐元素累加进行调整后，可以得到最终序列单模态的特征表示为：

$$\boldsymbol{U}^{k}=\mathrm{LN}\left(\mathrm{FC}\left(\boldsymbol{SA}^{k}\right)+\boldsymbol{SA}^{k}\right)\in\mathbf{R}^{2d_h\times(L-H+1)} \tag{5}$$

2.2 跨模态时序特征融合

特征融合是多模态学习的核心，因此在获得单模态特征表示后，本文基于多头注意力机制，挖掘两两模态组合之间双向的交互关系。传统的图文情感分析研究中，通常只会考虑从文本到图像的交互关系，将图像的信息附加到文本特征上，这是因为文本内容可以提供较为完整的信息，而视觉特征仅起到辅助增强情感表达效果。但对于本文的研究问题，三种模态的信息都是完整的，并且它们互为补充，共同传递演讲者的情感和态度，因此时序多模态的特征融合需要在时间尺度上发现双向的交互关系。

MulT 模型将基于多头注意力的特征融合方法推广到时序多模态问题上，利用模态 A 每个时刻的数据作为索引向量，计算与另一种模态 B 所有时刻数据的相似度，从而将模态 B 的信息附加到模态 A 中，实现了从模态 A 到模态 B(记作 $A\rightarrow B$)单方向的特征融合。这种特征融合方法能够处理不同长度的序列，在非对齐序列上也保留了较好的效果。本文同样是基于多头注意力机制，以文本 $\boldsymbol{U}^{T}$ 和语音 $\boldsymbol{U}^{A}$ 的跨模态融合为例，计算从文本到语音的融合时每个注意力头为：

$$Head_i\left(\boldsymbol{U}^T,\boldsymbol{U}^A\right)=\text{softmax}((\boldsymbol{W}^Q\boldsymbol{U}^T)^{\mathrm{T}}\cdot(\boldsymbol{W}^K\boldsymbol{U}^A)/\sqrt{2d_h/M})(\boldsymbol{W}^V\boldsymbol{U}^A)^{\mathrm{T}} \tag{6}$$

拼接所有注意力头输出的结果后，经过如式(4)和式(5)的前向映射和层归一化处理，得到单向融合结果 $\boldsymbol{CA}^{T\to A}\in\mathbf{R}^{2dh\times(L-H+1)}$。按照同样的方法可以获得从语音到文本的融合特征 $\boldsymbol{CA}^{A\to T}\in\mathbf{R}^{2dh\times(L-H+1)}$，拼接双向的融合结果获得完整的跨模态融合结果：

$$\boldsymbol{CA}^{TA}=\text{concat}\left(\boldsymbol{CA}^{T\to A},\boldsymbol{CA}^{A\to T}\right)\in\mathbf{R}^{4d_h\times(L-H+1)} \tag{7}$$

跨模态融合的特征依然保持时序特征，使用多头自注意力进一步的提取特征，发现序列自身的上下文信息，最终文本与语音信息融合后的结果记作 $\boldsymbol{F}^{TA}\in\mathbf{R}^{4dh\times(L-H+1)}$。此时三种单模态特征表示 $\boldsymbol{U}^T$, $\boldsymbol{U}^A$, $\boldsymbol{U}^V$ 和三种跨模态融合特征 $\boldsymbol{F}^{TA}$, $\boldsymbol{F}^{TV}$, $\boldsymbol{F}^{AV}$ 都是二维矩阵，为了方便下游任务模型的计算，本文使用平均池化整合所有时刻上的数据，并使用线性映射将单模态特征投影到与跨模态特征相同维度的空间。最终上游模型提取的六种特征表示共同拼接为完整的共享特征表示 $\boldsymbol{SF}=[\boldsymbol{SF}^T,\boldsymbol{SF}^A,\boldsymbol{SF}^V,\boldsymbol{SF}^{TA},\boldsymbol{SF}^{TV},\boldsymbol{SF}^{AV}]\in\mathbf{R}^{4dh\times6}$，输入到下游多任务学习模型中学习任务专属的融合特征。

2.3 多任务学习和任务专属特征融合

本文在下游模型中，添加情感极性分类和强度回归作为辅助任务，利用多任务学习的特点，帮助上游特征表示模型学习更具区分度和泛化性的特征，如图 5 所示。在 2.2 节中提取的特征表示 $\boldsymbol{SF}$ 在主任务和两项辅助任务之间共享，上游模型接受来自三项任务的梯度进行参数更新。三项任务之间使用硬参数共享机制，除输出层神经元数量和激活函数不同，其余结构全部统一。

Tian 等[11]从心理学和认知学角度设计的辅助任务具有可解释性，但是考虑到情感评分主任务是回归问题，而情感的强度通常是一个连续的实数值，不能简单地作为多分类问题。因此，根据情感评分回归主任务的样本标签 $y\in[-3,3]$，设置二分类任务检测情感极性 $y_P\in\{\text{positive, negative}\}$，同时设置回归任务预测情感强度 $y_I=\text{abs}(y)\in[0,3]$。

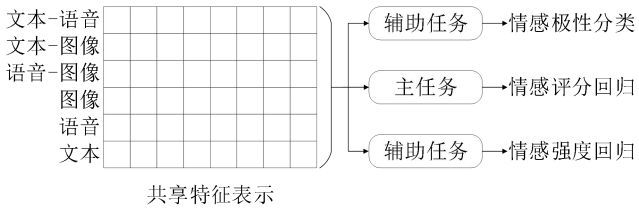


图 5 下游多任务学习框架

Fig. 5 Framework of downstream multi-task learning

文本、语音、图像这三种模态都能够传递一定的情感信息，但是在表现不同的情感时，它们的贡献度是变化的。此外，对于不同的任务目标，每种模态或特征的重要性也是不同的。将多任务学习集成到多模态情感分析中，需要根据任务的目标，衡量每种模态信息的重要性。

本文提出的任务专属特征融合方法如图 6 所示，在每项下游任务中设置独立的评分模块，根据任务目标学习每种特征表示的注意力权重。上游特征表示模型学习的共享特征由 6 种融合特征组成 $\boldsymbol{SF}=[\boldsymbol{SF}^T,\boldsymbol{SF}^A,\boldsymbol{SF}^V,\boldsymbol{SF}^{TA},\boldsymbol{SF}^{TV},\boldsymbol{SF}^{AV}]$，利用前向神经网络学习每种特征表示的注意力权重：

$$attn^k=\text{softmax}\left(\boldsymbol{W}_2^k\tanh\left(\boldsymbol{W}_1^k\boldsymbol{SF}^k+\boldsymbol{b}_1^k\right)+\boldsymbol{b}_2^k\right) \tag{8}$$

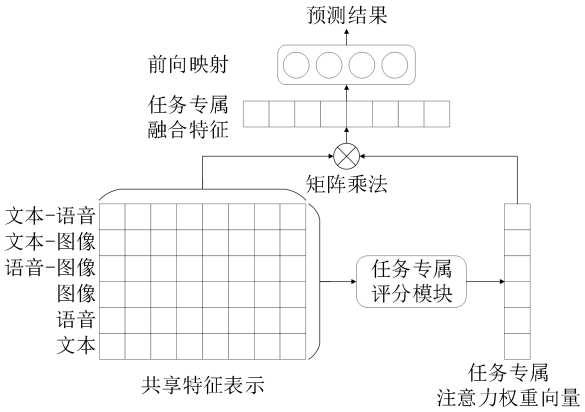


图 6 任务专属特征融合

Fig. 6 Task-specific feature fusion

其中，$\boldsymbol{W}_1^k$ 和 $\boldsymbol{W}_2^k$ 分别是两个全连接层的权重参数，$\boldsymbol{b}_1^k$ 和 $\boldsymbol{b}_2^k$ 是对应的偏置。每种特征表示对应的注意力权重构成了任务专属的注意力向量 $\boldsymbol{attn}=[attn^{TA},attn^{TV},attn^{AV},attn^{V},attn^{A},attn^{T}]$，与共享特征表示矩阵相乘获得任务专属的融合表示 $\boldsymbol{FF}_{task}=\boldsymbol{SF}\times\boldsymbol{attn}\in\mathbf{R}^{4dh}$。后续的预测则根据具体的任务目标，设置相应的输出层神经元数量和激活函数。本文所有的任务均设置输出层神经元数量为 1，主任务和强度回归任务不使用激活函数，而极性二分类任务使用 sigmoid 激活函数。回归任务使用平均绝对误差(Mean Absolute Error, MAE)作为损失函数，而分类任务使用交叉熵(Cross Entropy)作为损失函数。整个模型的训练基于三项任务的综合损失：

$$loss=\frac{\alpha\times loss_{main}+\beta\times loss_{polar}+\gamma\times loss_{inten}}{\alpha+\beta+\gamma} \tag{9}$$

其中α，β和γ为超参数，用于调节三项任务的训练程度。较大的参数值能够指导模型优先学习该项任务，加快该任务上的收敛速度。

3 实验

本节将通过定性和定量的实验校验改进模型的效果，所实验使用 Python 3.6.9 语言编写，使用深度学习框架 PyTorch 1.4.0 实现神经网络结构。实验环境为 Ubuntu18.04 系统，硬件设置为 Intel Core i9-9900K@3.6 GHz×16 处理器和 GeForce RTX 2080 显卡。

本文使用两个公开的多模态基准数据集 CMU-MOSI 和 CMU-MOSEI 对改进的模型进行评估。两个数据集中的样本都是由油管视频片段中分解出的文本、语音、图像三种时序模态构成，每个样本对应的情感标签为 $y\in[-3.0, 3.0]$范围内的实数值，即为情感评分。对于情感评分回归主任务，直接使用样本对应的情感评分作为回归目标。对于情感极性二分类任务，则将 $y\geq 0$ 的数据标记为积极状态，$y<0$ 则标记为消极状态。对于情感强度回归任务，则以情感评分的绝对值作为对应的预测目标。在计算七分类准确度时，则基于七级李克特量表情感表示模型，通过四舍五入将实数的情感评分映射为七个类别标签作为七分类的目标。

数据集中的文本部分使用预训练的 BERT(Bidirectional Encoder Representation from Transformers)模型进行编码，获得对应的嵌入表示作为文本特征。语音和图像部分则直接使用多模态开发工具包(CMU Multimodal SDK)[32]提供的特征。在对齐三种模态序列后，为了方便实验测试，本文通过截断和填充统一所有样本的序列长度，并按照指定编号划分数据集，相关统计信息如表 2 所示。

表 2 多模态基准数据集统计信息

Tab. 2 Statistics of multimodal benchmark datasets

统计信息	CMU-MOSI	CMU-MOSEI
训练集样本数量	1283	16315
验证集样本数量	229	1871
测试集样本数量	686	4654
文本特征维度	768	768
语音特征维度	74	74
图像特征维度	47	35
序列长度	20	40

3.1 定量实验

模型训练过程中选用 Adam 优化器，设置学习率为 5e-4，批训练样本数量为 128。上游特征表示模型中，使用 100 个高度为 3 的卷积核，BiGRU 隐藏神经元数量设置为 100。下游多任务学习模型中，设置 CMU-MOSEI 数据集上的超参数 α，β和γ分别为 1，1 和 1，设置 CMU-MOSI 数据集上的超参数全为 1。为了比较和评估本文提出的改进模型，选用以下几种多模态情感分析方法作为对比，实验结果如表 3 和表 4 所示，对比方法的结果全部引用自相应的原文献中。

RMFN(Recurrent Multistage Fusion Network)[33]：该模型将跨模态的融合过程分解为多个阶段进行，并使用循环神经网络捕获时序模态内部的信息。

MFM(Multimodal Factorization Model)：Tsai 等[34]提出了一种全新的视角来学习多模态特征表示，它能够将每种模态信息分解为共享的判别因子和独有的生成因子。

RAVEN(Recurrent Attended Variation Embedding Network)[35]：该方法基于注意力模型，使用非文本模态信息来调整词语的嵌入表示，它指出说话者的意图与非文本模态信息具有一定的关联，在理解人类语言时也需要考虑非文本的模态信息。

MCTN(Multimodal Cyclic Translation Network)[36]：该方法基于编码器和解码器结构，学习模态之间的转换关系，并利用循环一致性损失构建多模态特征表示。

MulT：该模型基于多头注意力机制和 Transformer 结构，学习模态两两之间的转换关系，能够捕捉跨模态的交互关系。

CIM-MTL：该方法是经典的基于多任务学习的多模态情感分析模型，它利用情感细粒度的多标签分类任务，辅助提升主任务的性能。

表 3 CMU-MOSEI 数据集上实验结果

Tab. 3 Experimental results on CMU-MOSEI dataset

模型	Acc-7/%	Acc-2/%	F1/%	MAE	Corr
MFM	45.0	76.9	77.0	0.710	0.540
RAVEN	50.0	79.1	79.5	0.614	0.662
MCTN	49.6	79.8	80.6	0.609	0.670
MulT	51.8	82.5	82.3	0.580	0.703
CIM-MTL	—	80.5	78.8	—	—
本文方法	51.9	84.7	84.7	0.582	0.707

表 4 CMU-MOSI 数据集上实验结果

Tab. 4 Experimental results on CMU-MOSI dataset

模型	Acc-7/%	Acc-2/%	F1/%	MAE	Corr
RMFN	38.3	78.4	78.0	0.922	0.681
MFM	36.2	78.1	78.1	0.951	0.662
RAVEN	33.2	78.0	76.6	0.915	0.691
MCTN	35.6	79.3	79.1	0.909	0.676
MulT	40.0	83.0	82.8	0.871	0.698
本文方法	36.0	81.2	81.2	0.948	0.639

考虑到主任务是情感评分回归任务，因此选用 MAE 和皮尔森相关系数(Pearson Correlation, Corr)为评价指标。此外，本文使用二分类准确度(Acc-2)，七分类准确度(Acc-7)和 F1 值作为分类性能的评价指标。根据表 3 和表 4 中的结果显示，本文的方法在 CMU-MOSEI 数据集上取得了最好的结果，而 MulT 模型在 CMU-MOSI 数据集上效果更好。MulT 模型在 CMU-MOSI 数据集上的结果优于在 CMU-MOSEI 数据集上的结果，而结合表 2 所示的统计信息可以发现，CMU-MOSEI 数据集的训练样本总量高于 CMU-MOSI 数据集。因此，可以得知 MulT 模型虽然在 CMU-MOSI 数据集上效果更好，但它在该数据集上过拟合，不能推广到 CMU-MOSEI 数据集上。而本文的方法在提供更多的训练样本后，其分类和回归表现均获得了提升，这也说明多任务学习能够有效地降低过拟合的风险，提升模型的泛化性。

表 5 CMU-MOSEI 数据集上的消融实验

Tab. 5 Ablation experiments on CMU-MOSEI dataset

单模态表示			跨模态融合	共享特征表示		多任务		评价指标		
CNN	BiGRU	自注意力	自注意力	单模态	跨模态	极性分类	强度回归	Acc-7/%	Acc-2/%	F1/%
×	√	√	√	√	√	√	√	51.1	84.2	84.2
√	×	√	√	√	√	√	√	51.0	84.2	84.3
√	√	×	√	√	√	√	√	51.1	83.9	84.1
√	√	√	×	√	√	√	√	51.5	84.1	84.1
√	√	√	√	×	√	√	√	49.9	83.3	83.6
√	√	√	√	√	×	√	√	50.4	84.3	84.4
√	√	√	√	√	√	×	√	51.1	83.6	83.6
√	√	√	√	√	√	√	×	49.5	84.2	84.3
√	√	√	√	√	√	√	√	51.9	84.7	84.7

为了进一步验证模型各部分模块的必要性和有效性，分别移除每一个模块，比较其对模型整体的影响。在CMU-MOSEI数据集上的实验结果如表5所示。根据二分类准确度和七分类准确度指标上的降低，可以衡量各部分模块对模型整体的影响。可以发现，当移除单模态特征表示部分的自注意力模块后，二分类准确度存在明显的降低。同样的，移除单模态共享特征也会影响到二分类准确度。而移除单模态和跨模态共享特征，都会造成七分类准确度的降低，也证明了多模态学习中，发现模态内部和模态之间信息的必要性。此外，在下游任务模型中，移除情感极性分类任务和情感强度回归任务，分别会对二分类和七分类准确度产生显著的影响，这个结果符合本文对情感极性和强度任务的定义，也证明了主任务能够通过相关任务的辅助而获得提升。

3.2 定性实验

本文在下游多任务学习部分中提出了任务专属特征融合方法，并在每个任务中添加专属的评分模块，按照特定的任务目标，计算相应的共享特征权重。为了理解不同任务目标与共享特征之间的对应关系，本文使用箱线图可视化每种共享特征表示对应的注意力权重，如图7所示。

箱线图能够展示一组数据的分布情况，从图7中的权重系数分布可以得出结论，对于不同的任务目标，每种共享特征对应的重要性也是不同的，这也证明了任务专属评分模块的必要性。如图7(a)所示，情感评分回归主任务的权重主要集中于文本-图像和文本-语音融合特征，以及文本单模态特征，这说明了情感评分任务对文本信息的依赖性。主任务对融合特征分配了较高的注意力权重，这也证明了利用相关的非文本信息，能够辅助增强文本的情感表达。而在图7(c)所展示的情感强度回归任务中，对文本-语音和文本-图像融合特征的依赖也证明了挖掘模态关联性的必要。在图7(b)中，情感极性分类任务的注意力权重则分散在三种融合特征和语音特征上。最后，通过可视化三种情感分析任务与六种共享特征的注意力权重，可以总结出以下三点结论，也进一步验证了1.2节中相关研究工作的结果：

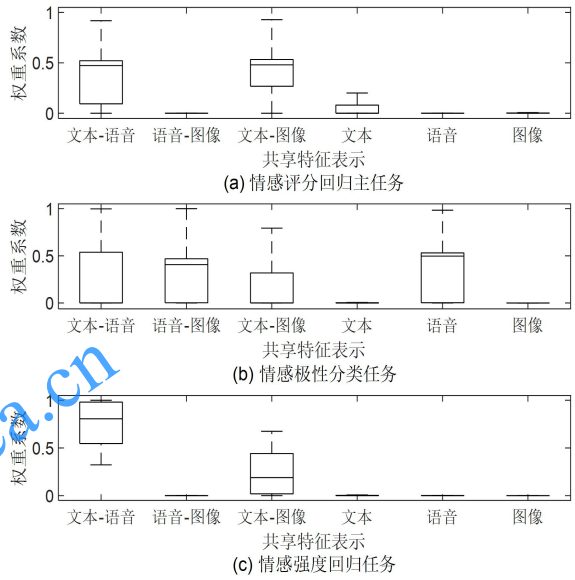


图7 权重系数可视化

Fig. 7 Visualization of weight coefficients

(1) 相较于语音和图像模态，文本模态通常蕴含更加丰富的情感信息，这也解释了早期情感分析工作大多集中于文本内容的原因；

(2) 利用非文本模态信息，能够更好的提取文本中的情感内容，增强情感表达能力，这说明了跨模态融合对情感分析工作的必要性和重要性；

(3) 不同的模态信息(包括单模态和跨模态)具有不同的重要性，并且会伴随具体的任务目标而变化，这说明多模态模型不能只关注于学习单模态和跨模态特征，也需要进一步考虑所提取的每种信息的重要性。

4 结语

多模态情感分析是情感计算领域新兴的研究重点，它不仅要求模型能够发现模态内部独有的特征，还要求能够正确捕捉模态之间的相互作用。而本文的研究对象是以油管视频为代表的多模态序列，这为多模态情感分析带来了新的问题。模态的时序特性要求模型能够充分挖掘单模态潜在的序列和

上下文信息，并且序列模态的相互作用是发生在时间尺度上。本文首先提出集成了卷积神经网络、双向门控循环神经网络和多头自注意力机制的单模态特征表示方法。卷积神经网络能够提取序列的局部特征，同时缩短序列长度并统一多模态序列的维度。双向门控网络能够挖掘前向和后向的序列信息，而多头自注意力则能够有效地提取上下文信息。其次，本文提出基于多头注意力的跨模态表征融合方法，挖掘两两模态之间，双向的交互关系，构建模态融合特征表示。最后，本文基于多任务学习思想，在下游模型中添加两项额外的辅助任务，利用任务之间的依赖关系，指导上游模型学习更具判别性和泛化性的特征表示。通过在两个经典多模态情感分析数据集上的实验，可以证明本文方法的有效性。

本文的方法依赖于多头注意力机制捕获模态自身与模态之间的信息，这种方法具有较高的计算复杂度和空间开销，而 MFM 模型给多模态学习指出了新的研究方向。共存的多模态之间具有共同的成分，也具有每种模态所独有的成分。通过对模态进行分解，能够更好的捕获模态的独有特征和共有信息，这种方法也具有更好地可解释性。因此，在未来的工作中，将针对模态分解方法展开进一步的深入研究。

参考文献

[1] YADOLLAHI A, SHAHRAKI A G, ZAIANE O R. Current state of text sentiment analysis from opinion to emotion mining[J]. ACM Computing Surveys, 2017, 50(2): 25:1-25:33.

[2] HONG M-S, JUNG J J. Multi-sided recommendation based on social tensor factorization[J]. Information Sciences, 2018, 447: 140-156.

[3] 蔡国永, 吕光瑞, 徐智. 基于层次化深度关联融合网络的社交媒体情感分类[J]. 计算机研究与发展, 2019, 56(6): 1312-1324. (CAI G Y, LV G R, XU Z. A hierarchical deep correlative fusion network for sentiment classification in social media [J]. Journal of Computer Research and Development, 2019, 56(6): 1312-1324.)

[4] TRUONG Q T, LAUW H W. VistaNet: visual aspect attention network for multimodal sentiment analysis[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 305-312.

[5] VERMA S, WANG C, ZHU L, et al. DeepCU: integrating both common and unique latent information for multimodal sentiment analysis[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. California: IJCAI, 2019: 3627-3634.

[6] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5634-5641.

[7] PHAM H, MANZINI T, LIANG P P, et al. Seq2Seq2Sentiment: multimodal sequence to sequence models for sentiment analysis[C]//Proceedings of the 1st Grand Challenge and Workshop on Human Multimodal Language. Stroudsburg, PA: Association for Computational Linguistics, 2018: 53-63.

[8] MAI S, HU H, XING S. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 164-172.

[9] TSAI Y-H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 6558-6569.

[10] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.

[11] TIAN L, LAI C, MOORE J. Polarity and intensity: the two aspects of sentiment analysis[C]//Proceedings of the 1st Grand Challenge and Workshop on Human Multimodal Language. Stroudsburg, PA: Association for Computational Linguistics, 2018: 40-47.

[12] AKHTAR M S, CHAUHAN D S, GHOSAL D, et al. Multi-task learning for multi-modal emotion recognition and sentiment analysis[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 370-379.

[13] ZHAO S, WANG S, SOLEYMANI M, et al. Affective computing for large-scale heterogeneous multimedia data: a survey[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(3s): 1-32.

[14] HOVY E H. What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis[M]//Language Production, Cognition, and the Lexicon. Berlin: Springer-Verlag, 2015: 13-24.

[15] MUNEZERO M, MONTERO C S, SUTINEN E, et al. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text[J]. IEEE Transactions on Affective Computing, 2014, 5(2): 101-111.

[16] BORTH D, JI R, CHEN T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]//Proceedings of the 13th ACM Multimedia Conference. Barcelona, New York: ACM, 2013: 223-232.

[17] GUILLAUMIN M, VERBEEK J J, SCHMID C. Multimodal semi-supervised learning for image classification[C]//Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2010: 902-909.

[18] BALTRUSAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[19] 陈郑淏, 冯翱, 何嘉. 基于一维卷积混合神经网络的文本情感分类[J]. 计算机应用, 2019, 39(7): 1936-1941. (CHEN Z H, FENG A, HE J. Text sentiment classification based on 1D convolutional hybrid neural network[J]. Journal of Computer Applications, 2019, 39(7): 1936-1941.)

[20] HUANG F, ZHANG X, ZHAO Z, et al. Image-text sentiment analysis via deep multimodal attentive fusion[J]. Knowledge-Based Systems, 2019, 167: 26-37.

[21] 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075-3080. (LI Y, DONG H B. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network[J]. Journal of Computer Applications, 2018, 38(11): 3075-3080.)

[22] CHEN F, JI R, SU J, et al. Predicting microblog sentiments via weakly supervised multimodal deep learning[J]. IEEE Transactions on Multimedia, 2018, 20(4): 997-1007.

[23] CHEN F, LUO Z, XU Y, et al. Complementary fusion of multi-features and multi-modalities in sentiment analysis[C]// Proceedings of the 3rd Workshop on Affective Content Analysis co-located with 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 2614: 82-99.

[24] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Stroudsburg, PA: Association for Computational Linguistics, 2017: 1103-1114.
[25] 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. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2247-2256.
[26] LIANG P P, LIU Z, TSAI Y-H H, et al. Learning representations from imperfect time series data via tensor rank regularization[C]// Proceedings of the 57th Conference of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 1569-1576.
[27] YU J, JIANG J. Adapting bert for target-oriented multimodal sentiment classification[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. California: IJCAI, 2019: 5408-5414.
[28] MAJUMDER N, HAZARIKA D, GELBUKH A F, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-Based Systems, 2018, 161: 124-133.
[29] CARUANA R. Multitask learning[J]. Machine Learning, 1997, 28(1): 41-75.
[30] 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.
[31] 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. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2236-2246.
[32] ZADEH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5642-5649.
[33] LIANG P P, LIU Z, ZADEH A, et al. Multimodal language analysis with recurrent multistage fusion[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2018: 150-161.
[34] TSAI Y-H H, LIANG P P, ZADEH A, et al. Learning factorized multimodal representations[C]//Proceedings of the 7th International Conference on Learning Representations. La Jolla, CA: ICLR, 2019.
[35] WANG Y, SHEN Y, LIU Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 7216-7223.
[36] 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 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6892-6899.

This work is partially supported by the National Natural Science Foundation of China (61772282).

ZHANG Sun, born in 1994, PhD candidate. His research interests include deep learning, sentiment analysis and text classification.

YIN Chunyong, born in 1977, Ph. D., professor， Ph. D. supervisor. His research interests include cyberspace security, big data mining and privacy protection, artificial intelligence and new computing.