计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：面向多模态情感分析的双模态交互注意力
作者：包广斌，李港乐，王国雄
网络首发日期：2021-08-05
引用格式：包广斌，李港乐，王国雄．面向多模态情感分析的双模态交互注意力．计算机科学与探索．
https://kns.cnki.net/kcms/detail/11.5602.TP.20210804.1715.003.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2105071

面向多模态情感分析的双模态交互注意力

包广斌，李港乐+，王国雄
兰州理工大学 计算机与通信学院，兰州 730050
+ 通信作者 E-mail: 1450316716@qq.com

摘 要:针对现有多模态情感分析方法中存在情感分类准确率不高、难以有效融合多模态特征等问题，通过研究分析相邻话语之间的依赖关系和文本、语音和视频模态之间的交互作用，建立一种融合上下文和双模态交互注意力的多模态情感分析模型。该模型首先采用双向门控循环单元(Bidirectional Gated Recurrent Unit，BiGRU)捕获各模态中话语之间的相互依赖关系，得到各模态的上下文信息。为了学习不同模态之间的交互信息，提出了一种双模态交互注意力机制来融合两种模态的信息，并将其作为条件向量来区分各模态信息对于情感分类的重要程度；然后结合自注意力、全连接层组成多模态特征融合模块，挖掘模态内部和模态之间的关联性，获得跨模态联合特征。最后，将得到的上下文特征和跨模态联合特征进行拼接，经过一层全连接层后馈送至 Softmax 进行最终的情感分类。在公开的多模态情感分析数据集CMU-MOSI(CMU Multimodal opinion-level Sentiment Intensity)上对所提出的模型进行评估，实验结果表明，相比现有模型，该模型在多模态情感分类任务上的表现是有效的和先进的。

关键词:多模态；情感分析；双向门控循环单元；上下文；双模态交互注意力；特征融合
文献标志码：A **中图分类号：**TP391

Bimodal interactive attention for multimodal sentiment analysis

BAO Guangbin, LI Gangle+, WANG Guoxiong
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

Abstract: Aiming at the problems of low accuracy of sentiment classification and difficulty in effectively fusing multimodal features in existing multimodal sentiment analysis methods, a multimodal sentiment analysis model combines context and bimodal interactive attention is established by analyzing the dependence between adjacent utterances and the interaction among text, audio and video modalities. Firstly, the model adopts a bidirectional gated recurrent unit (Bidirectional Gated Recurrent Unit, BiGRU) to capture the dependence between utterances in each modal, and the context information of each modal is obtained. In order to learn the interactive information between different modalities, a bimodal interactive attention mechanism is proposed to fuse the information of the two modalities and use it as a condition vector to distinguish the importance of each modal for sentiment classification. Then, combining self-attention and fully connected layers to form a multimodal feature fusion module, mining the correlation of information within and between modalities, and obtaining cross-modal joint features. Finally, the ob-

基金项目：国家自然科学基金（51668043）；甘肃省自然科学基金（18JR3RA156）。
This work was supported by the National Natural Science Foundation of China (51668043), the Natural Science Foundation of Gansu Province (18JR3RA156).

tained contextual features and cross-modal joint features are spliced, and then fed to Softmax for the final sentiment classification after a fully connected layer. The proposed model is evaluated on the public multimodal sentiment analysis dataset CMU-MOSI (CMU Multimodal opinion-level Sentiment Intensity). The experimental results show that compared with the existing models, the performance of the model on multimodal sentiment classification task is effective and advanced.

Key words: multimodal; sentiment analysis; bidirectional gated recurrent unit; context; bimodal interactive attention; feature fusion

随着移动互联网和和社交媒体的蓬勃发展，越来越多的用户通过 YouTube、微博、抖音等社交媒体讨论时事、表达观点、分享日常等，从而产生了海量的具有情感取向的多模态数据。在社交媒体平台上，用户上传的视频是多模态数据的重要来源之一[1]。视频数据通常包含三种模态：即描述用户观点的文本、表达用户面部表情的图像以及记录用户语音语调的音频。针对这些多模态数据进行情感分析将有利于了解人们对某些事件或商品的观点和态度，在舆情分析、心理健康、政治选举等方面都有着巨大的应用价值[2]。

与传统的单模态情感研究相比，多模态情感分析的目标是通过融合多个模态的数据来推断目标序列的情感状态[3]。如图 1 显示了文本、面部表情和语音语调对于情感分类的作用。其中，视频中说话人关于某部电影发表评论：“The only actor who can really sell their lines is Erin.”这条评论是一个陈述句，而且没有明显体现情感取向的词语，所以仅仅依据这句话所传达的信息很难判断出说话人的情感状态，但如果为这句评论加入说话人的面部表情(Looks Sad)和语音语调(Loud voice)，则可以反映出说话人目前的情感状态是消极的。因此，对于多模态情感分析任务，文本、语音和视频模态之间的语义和情感关联能够为情感分类带来重要的补充信息。

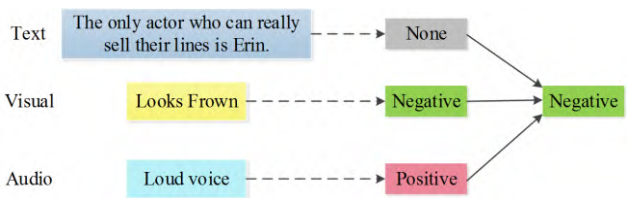

Fig. 1 The effect of text, facial expressions and voice intonation on sentiment classification

图 1 文本、面部表情和语音语调对于情感分类的作用

由于文本、语音和视频特征在时间、语义维度上存在较大差异，所以目前大多数多模态情感分析方法是将所有可用的模态特征直接映射到一个共享空间中[4]，学习不同模态之间复杂的交互作用。但是，大多数情况下，并不是融合的模态信息越丰富，情感分类的准确率就越高，这主要是因为不同模态的信息对于情感分类的贡献是不相等的[5]。

为了解决上述问题，本文提出了一种融合上下文和双模态交互注意力的多模态情感分析方法，该方法首先采用 BiGRU[6]分别捕获文本、语音和视频序列的上下文特征。然后利用不同模态之间存在的语义和情感关联，设计了一种双模态交互注意力，并结合自注意力[7]和全连接层构造了一个层次化的多模态特征融合模块，旨在通过注意力机制更多地关注目标序列及其上下文信息与各模态之间的相关性，从而帮助模型区分的哪些模态信息对于判别目标序列的情感分类更加重要，实现跨模态交互信息的有效融合。最后，在 CMU-MOSI[8]数据集上进行实验，实验结果表明，相比现有的多模态情感分类模型，该模型在准确率和 F1 分数上均有所提升。

1 相关工作

多模态情感分析主要致力于联合文本、图像、语音与视频模态的情感信息来进行情感的识别与分类，是自然语言处理、计算机视觉和语音识别相交叉的一个新兴领域[9]。与单一模态的情感分析相比，多模态情感分析不仅要学习单模态的独立特征，还要融合多种模态的数据[10]。多模态融合主要是通过建立能够分析和处理不同模态数据的模型来为情感分类提供更多的有效信息。Zadeh 等人[11]利用模态之间的联系建立了一种张量融合网络模

型，采用三倍笛卡尔积以端到端的方式学习模态之间的动力学。Zadeh 等人[12]提出了一种可解释的动态融合图 DFG (Dynamic Fusion Graph)模型，用于研究跨模态动力学的本质，并根据每个模态的重要性动态改变其结构，从而选择更加合理的融合图网络。Chen 等人[9]提出利用时间注意力的门控多模态嵌入式模型来实现多模态输入时单词级别的特征融合，该方法有效地缓解了噪声对特征融合的影响。上述方法在进行特征提取时都将每个话语看作独立的个体，忽略了与上下文之间的依赖关系。

多模态情感分析的研究数据通常来自社交网站上用户上传的视频，这些视频数据被人为的划分成视频片段序列，而片段序列之间往往存在着一定的语义和情感联系。所以，当模型对目标序列进行情感分类时，不同片段序列之间的上下文可以提供重要的提示信息。Poria 等人[13]建立了一种基于 LSTM(Long Short-Term Memory)的层次模型来捕捉视频片段间的上下文信息。Majumder 等人[14] 通过保持两个独立的门控循环单元来跟踪视频中对话者的状态，有效地利用了说话者之间的区别和对话中的上下文信息。Shenoy 等人[15]提出的基于上下文感知的 RNN（Recurrent Neural Network）模型能够有效地利用和捕获所有模态对话的上下文用于多模态情绪识别和情感分析。Kim 等人[16]建立了一种基于多头注意力的循环神经网络模型，该模型采用 BiGRU 和注意力机制来捕获会话的上下文信息的关键部分。但是，现在人们表达情感的方式已不再局限于单一的文字，往往通过文本、图像、视频等多种模态相结合的方式共同传递信息，那么如何有效利用多模态信息进行情感分析仍是一项艰巨的任务。

近年来，注意力机制已被广泛应用于 NLP 领域。研究表明，注意力机制能够聚焦于输入序列的关键信息，并忽略其中不相关的信息，从而提高模型的整体性能。因此，越来越多的研究人员尝试将注意力机制应用于探索模态内部和不同模态之间的交互作用。Zadeh 等人[17]提出了一种多注意力循环神经网络 MARN(Multi-attention Recurrent Network)，利用多注意力模块 MAB(Multi-attention Block)发现模态之间的相互作用，并将其存储在长短时混合记忆 LSTHM(Long-short Term Hybrid Memory)的循环网络中。Chen 等人[18]提出利用多头交互注意力来学习文本、语音和视频模态之间的相关性。Sunny 等人[19]了一种高阶通用网络模型来封装模态之间的时间粒度，从而在异步序列中提取信息，并利用 LSTM 和基于张量的卷积神经网络来发现模态内部和模态之间的动力学。

综上所述，随着深度学习研究的不断深入，多模态情感分析实现了跨越式的进步和发展，但如何有效地利用单模态独立特征和多模态交互特征进行建模依旧是多模态情感分析所面临的主要问题。为此，本文在现有多模态情感分析方法的基础上，提出了一种融合上下文和双模态交互注意力的多模态情感分析模型，旨在利用 BiGRU 和注意力机制更多地关注相邻话语之间的依赖关系以及文本、语音和视频模态之间的交互信息并为其分配合理的权重，实现多模态特征的有效融合，从而提高多模态情感分类的准确率。

2 融合上下文和双模态交互注意力的模型

本文针对现有多模态情感分析方法中存在情感分类准确率不高、难以有效融合多模态特征等问题，提出了一种融合上下文和双模态交互注意力的多模态情感分析模型(Multimodal sentiment analysis model based on context and bimodal interactive attention)，简称 Con-BIAM 模型，如图 2 所示。具体来说，Con-BIAM 模型分为以下四个部分：

（1）针对文本、语音和视频模态数据的不同特点，构建不同的神经网络提取单模态特征；

（2）利用 BiGRU 分别编码文本、语音和视频序列，然后将其映射到共享的语义空间中，在每个模态的不同时间步长上捕获视频目标序列的上下文信息；

（3）利用不同模态之间的交互作用，设计了一种新颖的双模态交互注意力机制融合不同模态的信息；然后通过双模态交互注意力、自注意力和全连接层构造多模态特征融合模块，得到跨模态联合

特征；

（4）将得到的上下文特征和跨模态联合特征连接起来，经过一层全连接层后馈送至 Softmax 进行最终的情感分类。

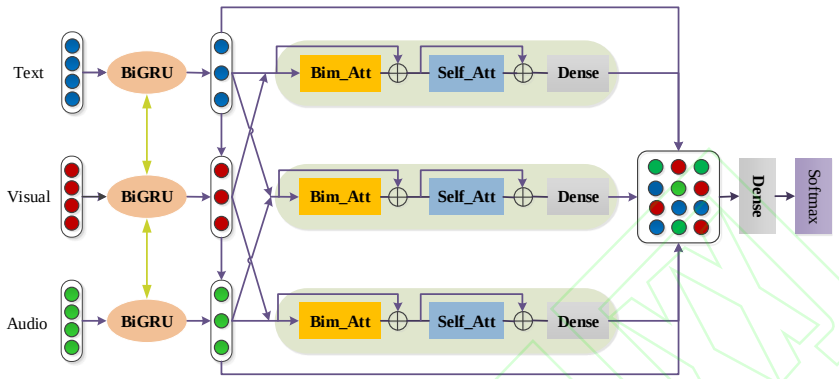

Fig.2 Model structure combining context and bimodal interactive attention

图 2 融合上下文和双模态交互注意力的模型结构

2.1 特征提取

为了获取视频中的文本、语音和视觉特征，采用卡内基梅隆大学提供的多模态数据分析工具 CMU-Multimodal Data SDK[17]进行提取。对于文本数据，首先将视频中的每个话语进行转录，然后将其表示为 Glove 词向量，输入至卷积神经网络中提取文本特征。为了有效地利用视频中的动态信息，使用 3D-CNN[20](3D Convolutional Neural Networks)从视频中提取视觉特征。在实验过程中，32 个特征图(f_m)和 $5\times5\times5$ ($f_d\times f_h\times f_w$)的过滤器取得了最优的结果。对于音频模态数据，利用 openSMILE[21]工具包以 30Hz 的帧速率和 100ms 的滑动窗口提取音频特征。

通过上述方法，原始视频片段序列将被表示为包含三种模态的特征向量 D，即 $D=[D_t,D_a,D_v]$。假设单模态特征的维度为 u，文本特征可表示为 $D_t\in\mathbb{R}^{u\times d_t}$；视觉特征可表示为 $D_a\in\mathbb{R}^{u\times d_v}$；语音特征可表示为 $D_v\in\mathbb{R}^{u\times d_a}$。

2.2 上下文特征表示

本文将预处理后的文本(D_t)、语音(D_a)和视频(D_v)特征分别输入至 BiGRU 中提取序列的上下文信息。考虑到不同模态数据的异构性，利用 Dense 层在时间维度上提取目标序列与上下文特征之间的长跨度信息，获得相同数据维度的上下文特征表示。

假设数据集包含 N 个视频片段，每个视频片段对应一个固定情感强度的观点。视频中包含的一系列片段序列可表示为：

$$X_i=\left[x_{i,1},x_{i,2},\ldots,x_{i,U_i}\right] \tag{1}$$

其中，X_{i,U_i} 表示第 i 个视频中的第 U_i 个片段序列；U_i 表示视频中所包含话语的最大长度。

此外，为了更加准确地对视频片段 X_i 进行情感分类，将 $X_{i,k}$ 定义为 X_i 的上下文：

$$X_{i,k}=[X_{i,k}\mid\forall k\le U_i,k\ne i] \tag{2}$$

其中，k 表示视频中其他片段序列的长度。

BiGRU 由两个方向相反的 GRU(Gated Recurrent Unit)构成，能够有效地捕获序列中上下文的长依赖关系，解决 RNN 训练过程中出现的梯度消失和梯度爆炸问题[13]。在 BiGRU 中，正向和反向输入的特征向量会得到对应时刻的隐藏层表示，之后通过拼接操作得到具有上下文信息的文本、视觉和语音特征。双向门控循环单元的结构如图 3 所示。

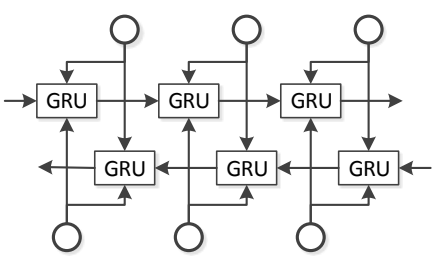

Fig.3 BiGRU structure model diagram

图 3 BiGRU 结构模型图

每个 GRU 单元的工作原理如下：

$$r_t=\sigma\left(W_r x_t+U_r h_{t-1}\right) \quad (3)$$

$$z_t=\sigma\left(W_z x_t+U_z h_{t-1}\right) \quad (4)$$

$$\tilde{h}_t=tanh\left(r_t * U h_{t-1}+W x_t\right) \quad (5)$$

$$h_t=(1-z_t) * h_{t-1}+z_t * \tilde{h}_t \quad (6)$$

其中，X_t 是当前节点的输入序列，h_{t-1} 是上一个 GRU 单元传输下来的状态，r_t 是 GRU 的重置门，z_t 是 GRU 的更新门，$W_r, W_z, U_r, U_z \in R^{d \times d}$ 是训练过程中要学习的参数，σ 是 Sigmoid 函数，$*$ 表示对应元素相乘。

为了深度挖掘单模态特征的内部相关性，将得到的具有上下文信息的单模态特征分别映射到各自的语义空间中。计算过程如下：

$$H_T=\tanh(W_t B_T+b_T) \quad (7)$$

$$H_A=\tanh(W_t B_A+b_A) \quad (8)$$

$$H_V=\tanh(W_t B_V+b_V) \quad (9)$$

其中，W_t, b_T, b_A, b_V 分别是激活函数 tanh 的参数，B_T, B_A, B_V 是经过 BiGRU 得到的文本、语音和视觉特征。$H_T \in \mathbb{R}^{u \times d}, H_A \in \mathbb{R}^{u \times d}, H_V \in \mathbb{R}^{u \times d}$ 分别表示最终输出的具有上下文信息的文本、语音和视觉特征向量，d 表示 Dense 层中神经元的数量。

2.3 特征融合模块

对于多模态情感分析任务，不同模态的数据包含了各自的情感信息，它们彼此不同却又相辅相成。因此，在基于模态内部关系建模的同时关注另一种模态信息的补充作用，能够有效地提升模型的分类性能。此外，在进行多模态信息融合时，不同模态的信息对情感分类结果的重要性也是不同的。因此，对多模态信息进行建模时，需要有选择性地区分各模态信息对目标序列的情感预测的重要程度，增强重要信息所占的权重，从而输出更有效的跨模态联合特征表示。

由此，本文提出了一种多模态特征融合模块（Multimodal feature fusion module），简称 MFM 模块。该模块采用层次化的融合策略融合所有的模态特征，主要由两层注意力机制和一个全连接层串联构成。首先第一层是双模态交互注意力(Bimodal Interactive Attention)层，简称 Bim_Att，Bim_Att 能够将两种模态的融合特征作为条件向量，强化与模态间重要交互特征的关联，弱化与次要交互特征的关联，深度探索不同模态之间的交互性；第二层是自注意力层(Self Attention)，简称 Self_Att，用于捕获目标序列及其上下文信息与模态自身的相关性，从而减少对外部信息的依赖；最后一层是全连接层，用于提取双模态交互融合信息和单模态内部信息，输出跨模态联合特征。

为了进一步增强模态之间的交互性，本文提出了一种双模态交互注意力机制，整体结构如图 4 所示。双模态交互注意力机制类似于一种门控机制，能够将文本、语音和视觉特征进行两两融合，即文本+视频，文本+语音和语音+视频，并有条件地计算不同模态之间的交互向量。以文本(H_T)和语音(H_A)为例，首先将两种模态的信息进行拼接，并经过一层全连接层捕获模态之间的交互信息，得到双模态联合特征 F_1；接着在激活函数 Sigmoid 的作用下生成条件向量 S_1，用于约束每个模态内部的相似程度，增加强关联特征所持的比重。计算过程如公式(10)-(11)所示。

$$F_1=\tanh(W_1(H_A \oplus H_T)+b_1) \quad (10)$$

$$S_1=Sigmoid(F_1) \quad (11)$$

其中，$\oplus$ 表示向量的拼接操作，W_1 表示随机初始化的权重矩阵，b_1 表示偏置项。

为了对不同模态的向量进行不同程度的关注，利用条件向量 S_1 分别与 H_T 和 H_A 结合得到文本条件向量 N_1 和语音条件向量 N_2，接着将两种模态的条件向量 N_1 和 N_2 进行矩阵乘法计算，得到跨模态联合矩阵 O_1；然后使用 Softmax 函数计算跨模态联合矩阵的概率分布 a_1，最后，将之前得到的双模态联合特征 F_1 乘以概率分布 a_1，以提升权重的方式来强化关键信息的比重，得到双模态交互注意力矩阵 Bim_{Att}。计算过程如公式(12)-(16)所示。

$$N_1=H_A \odot S_1 \quad (12)$$

$$N_2=H_T \odot S_1 \quad (13)$$

$$O_1=N_1 \bullet N_2 \quad (14)$$

$$a_1=\frac{e^{O_1(i,j)}}{\sum_{k=1}^{u}e^{O_1(i,k)}} \quad i,j=1,2,...,u \qquad (15)$$

$$Bim_{Att_1}=a_1 \bullet F_1 \qquad (16)$$

其中，$\odot$ 表示对应元素相乘，$\bullet$ 表示矩阵乘法。

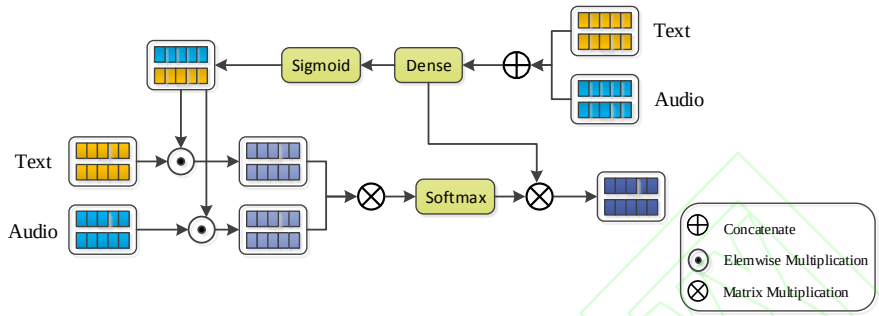

Fig.4 Structure diagram of bimodal interactive attention (Bim_Att)

图 4 双模态交互注意力（Bim_Att）结构图

经过双模态交互注意力输出的特征输入至自注意力层，获得模态内部相关特征 $Self_{Att}$。以文本模态为例，计算过程如公式(17)-(19)所示。

$$C_2=Bim_{Att_1}\oplus H_T\oplus H_A \qquad (17)$$

$$a_2=\frac{e^{C_2(i,j)}}{\sum_{k=1}^{u}e^{C_2(i,k)}} \quad i,\ j=1,2,.\iota \qquad (18)$$

$$Self_{Att_1}=a_2 \bullet C_2 \qquad (19)$$

其中，C_2 表示文本特征、语音特征与双模态联合特征拼接得到的特征向量，$Self_{Att}$ 表示信息过滤后的注意力特征向量。

最后，将得到注意力特征向量与上下文特征向量进行拼接，并使用全连接层整合得到的模态间交互特征和模态内部特征，并输入至 Softmax 进行情感分类，其计算过程如下。

$$C_3=Bim_{Att_1}\oplus Bim_{Att_2}\oplus Bim_{Att_3}\oplus H_T\oplus H_A\oplus H_v \qquad (20)$$

$$F_2=\mathrm{ReLU}(W_R C_3+b_R) \qquad (21)$$

$$Con\_BIAM=Soft\max(F_2) \qquad (22)$$

其中，$F_2\in\mathbb{R}^{u\times d'}$，$d'$ 表示全连接层输出的特征维度，W_R 和 b_R 是激活函数 Relu 的权重和偏置。

3 实验与结果分析

3.1 数据集

本文使用多模态情感分析数据集 CMU-MOSI 进行实验，简称 MOSI。该数据集由 89 位不同英语演讲者对来自 YouTube 网站中的主题进行评论，共有 93 个视频。数据集中共包含 3702 个观点片段，共计 26295 个单词。每个视频片段的情感强度在 $y\in[-3.0,3.0]$ 的线性范围内，其中大于或等于 0 的情感值表示正面情绪，小于 0 的情感值表示负面情绪。本实验将数据集划分为训练集，验证集和测试集，分别设置为 52，10，31。每个集合分别包含 1151、296 和 752 个视频片段。

3.2 实验设置

本实验所有代码都是在 Pycharm 代码编辑器上采用 Tensorflow 和 Keras 深度学习框架编写，利用显存为 32GB 的 GPU(NVIDIA Tesla V100)进行模型的训练。实验参数设置如表 1 所示。

Table 1 Experimental parameter settings

表 1 实验参数设置

参数	值
词向量维度	50
BiGRU 隐藏单元	300
全连接层神经元	100
Dropout	0.5
学习率	0.001
批处理	32
迭代次数	30
优化函数	Adam
损失函数	Categorical_crossentropy

本文选取 F1 分数和准确率(Accuracy)作为分类性能的评价指标。F1 分数和 Accuracy 的值越大，说明模型的整体性能越好。为了进一步验证模型的

有效性，将本文提出的 Con-BIAM 模型与现有的一些多模态情感分析模型进行对比，实验结果如表 2 所示。

3.3 实验结果分析

表 2 列出了不同模型在 MOSI 数据集上的实验结果。图 5 是 Con-BIAM 模型在 MOSI 数据集上得到的混淆矩阵。

Table 2　Experimental results on MOSI

表 2　在 MOSI 上的实验结果

Model	Accuracy	F1 score
GME-LSTM[9]	76.50	73.40
MARN[17]	77.10	77.00
TFN[11]	77.10	77.90
DialogueRNN[14]	79.80	79.48
BC-LSTM[13]	80.30	-
Multilogue-Net[15]	81.19	80.10
Con-BIAM	81.91	85.40

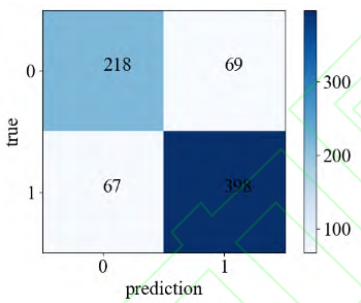

Fig.5　Con-BIAM model confusion matrix on MOSI dataset

图 5　Con-BIAM 模型在 MOSI 数据集上的混淆矩阵

实验结果表明，本文提出的 Con-BIAM 模型在准确率和 F1 分数这两个评价指标上的表现都要优于其他对比模型，准确率和 F1 分数分别提升了 5.41% 和 12%，尤其是对比现有先进的 Multilogue-Net 模型，准确率提升了 0.72%，F1 提升了 5.3%。这充分地说明了融合上下文和双模态交互注意力的多模态情感分析(Con-BIAM)在多模态情感分类任务上的有效性和先进性。此外，根据上述实验结果可以看出，Con-BIAM 模型的 F1 值与其他模型相比具有较大提升，这可能是因为不同层次不同组合的模态融合方法关注到了模态的内部信息和更高层次的模态交互信息，使得模型的精确率和召回率分别达到了 85.22%和 85.59%，进而增大了模型的 F1 值，提高了模型的分类性能。

4 对比实验

为了进一步分析模态之间的联合特征对模型最终分类效果的贡献程度，在 MOSI 数据集上分别针对双模态和三模态联合特征，选择以下几种多模态情感分析方法进行对比，实验结果如表 3 和表 4 所示。

Table 3　Accuracy of different models in bimodal and trimodal feature fusion

表 3 不同模型在双模态、三模态特征融合的准确率

Metric	Dialogue-RNN[14]	Multilogue-Net[15]	Con_BIAM
T+A	79.80	80.18	80.45
V+T	78.90	80.06	80.98
A+V	73.90	75.16	63.96
A+V+T	79.80	81.19	81.91

Table 4　F1 scores of different models in bimodal and trimodal feature fusion

表 4　不同模型在双模态、三模态特征融合的 F1 分数

Metric	Dialogue-RNN[14]	Multilogue-Net[15]	Con_BIAM
T+A	78.32	79.88	84.14
V+T	78.12	79.84	84.43
A+V	73.92	74.04	75.20
A+V+T	79.48	80.10	85.40

实验结果表明，与其他模型相比，除了语音和视频模态的融合之外，Con-BIAM 模型的其他模态融合方式都达到了最好的结果。其中，三种模态(文本、语音和视觉)融合的分类效果最佳，证明了多模态信息的必要性。在双模态融合的实验中，文本+图像和文本+语音融合分类准确率高于语音+视频的融合。这一方面说明了文本模态的情感特性更为显著；另一方面也反应了语音和视频模态的情感特性较弱，可能存在噪声的干扰。

为了进一步分析视频片段的上下文信息、自注意力和双模态交互注意力对模型性能的贡献，本文设计了三组对比实验，比较不同模块对于模型整体性能的影响。在 MOSI 数据集上对比实验的结果如图 6 所示。

（1）Con_BIAM(GRU)：使用 GRU 代替模型中 BiGRU，比较上下文信息对模型性能的影响。

（2）Con_BIAM(Self_Att)：舍弃双模态交互

注意力机制，保留自注意力机制，探究两种模态之间的交互信息对分类效果的影响。

（3）Con_BIAM(Bim_Att)：舍弃自注意力机制，保留双模态交互注意力，探究单模态情感信息对分类效果的影响。

（4）Con_BIAM：本文所提出模型。

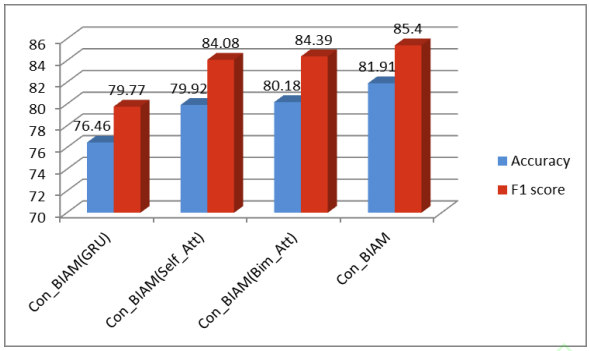

Fig.6 Comparative experiment on MOSI dataset

图 6 在 MOSI 数据集上的对比试验

实验结果表明，对于 MOSI 数据集，舍弃 Con_BIAM 模型中的任一重要模块，都会使得模型的分类性能下降。首先，相比于 GRU 模型，基于 BiGRU 的模型准确率提升了 2.52%，说明了对于视频中某一片段序列，序列前面和后面的视频片段都会对它产生一定的影响，而 BiGRU 能够同时捕捉到视频片段序列前向和反向的信息。其次，多模态特征融合模块中的双模态交互注意力和自注意力对情感分类的准确率分别贡献了 1.2%和 0.94%，F1 值也分别提升了 2.67%和 2.36%。这主要是因为文本、语音和视频模态内部与模态之间存在着大量的情感信息，而本文所设计的多模态特征融合模块能够同时提取单模态信息和双模态融合信息，并通过注意力机制有选择地关注有利于情感分类的模态信息，从而提高了模型分类性能。

5 结论

本文建立了一种融合上下文和双模态交互注意力的多模态情感分析模型，利用视频片段的上下文信息和不同模态之间的交互信息来预测情感分类。该模型首先采用 BiGRU 捕获文本、语音和视频序列之间的上下文信息。然后，通过双模态交互注意力、自注意力和全连接层构成的多模态特征融合模块，关注目标序列及其上下文信息与模态内部和模态之间的关联性，实现了多模态信息的有效融合。最后，将得到的上下文特征和跨模态联合特征输入至分类器进行情感分类。在 MOSI 数据集上的实验结果证明了所提出的模型在多模态情感分类任务上的有效性和优异性。在未来的工作中，将针对多模态融合过程中所出现的语义冲突和噪声问题展开进一步研究。

参考文献：

[1] GHOSAL D, AKHTAR M S, CHAUHAN D, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Oct 31-Nov 4, 2018. Stroudsburg, USA: ACL, 2018: 3454-3466.

[2] LIN M H, MENG Z Q et al. Multimodal Sentiment Analysis Based on Attention Neural Network[J]. Computer Science, 2020, 47(S2): 508-514+548.

林敏鸿，蒙祖强. 基于注意力神经网络的多模态情感分析[J]. 计算机科学, 2020, 47(S2): 508-514+548.

[3] LIU J, ZHANG P, LIU Y, et al. Summary of Multi-modal Sentiment Analysis Technology[J/OL]. Journal of Frontiers of Computer Science and Technology: 1-19 [2021-05-18]. http://kns.cnki.net/kcms/detail/11.5602.TP.20210324.1400.016.html.

刘继明，张培翔，刘颖，等. 多模态的情感分析技术综述[J/OL].计算机科学与探索: 1-19 [2021-05-18]. http://kns.cnki.net/kcms/detail/11.5602.TP.20210324.1400.016.html.

[4] HE J, ZHANG C Q, LI X Z, ET AL. Survey of research on multimodal fusion technology for deep learning[J].Computer Engineering, 2020, 46(5) : 1-11.

何俊，张彩庆，李小珍，等. 面向深度学习的多模态融合技术研究综述[J]. 计算机工程, 2020, 46(5) : 1-11.

[5] PORIA S, CAMBRIA E, HAZARIKA D, et al. Multi-level multiple attentions for contextual multimodal sentiment

analysis[C]//Proceedings of IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, Nov 18-21, 2017. Washington D.C., USA: IEEE, 2017: 1033-1038.

[6] KUMAR A, VEPA J. Gated mechanism for attention based multi modal sentiment analysis[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, May 4-8, 2020. Washington D.C., USA: IEEE, 2020: 4477-4481.

[7] LIN Z, FENG M, SANTOS C N, et al. A structured self-attentive sentence embedding[EB/OL]. (2017-03-09)[2021-05-03]. http://arxiv.org/pdf/1703.03130.pdf.

[8] 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.

[9] CHEN M, WANG S, LIANG P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning[C]//Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, Scotland, Nov 13-17th 2017. New York, USA: ACM, 2017: 163-171.

[10] ZHANG Y Z, RONG L, SONG D W, et al. A Survey on Multimodal Sentiment Analysis. Pattern Recognition and Artificial Intelligence, 2020, 33(5): 426-438.

张亚洲，戎璐，宋大为，等. 多模态情感分析研究综述[J]. 模式识别与人工智能, 2020, 33(5): 426-438.

[11] 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, Copenhagen, Denmark, Sep 9-11, 2017. Washington D.C., USA: IEEE, 2017: 1103-1114.

[12] 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, Jul 15-20, 2018. Stroudsburg, USA: ACL, 2018: 2236-2246.

[13] 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, Jul 30-Aug 4th, 2017. Stroudsburg, USA: ACL, 2017: 873-883.

[14] MAJUMDER N, PORIA S, HAZARIKA D, et al. Dialoguernn: An attentive rnn for emotion detection in conversations[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, Jan 27–Feb 1, 2019. Palo Alto, USA: AAAI, 2019, 33(01): 6818-6825.

[15] SHENOY A, SARDANA A. Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation[J]. ACL, 2020: 19-28.

[16] KIM T, LEE B. Multi-Attention Multimodal Sentiment Analysis[C]//Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, June 8-11, 2020. New York, USA: ACM, 2020: 436-441.

[17] 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, LA, USA, Feb 2-7, 2018. Palo Alto, USA: AAAI, 2018: 5642.

[18] XI C, LU G, YAN J. Multimodal sentiment analysis based on multi-head attention mechanism[C]//Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Stockholm, Sweden, Oct 10-Nov 14, 2020. New York, USA: ACM, 2020: 34-39.

[19] VERMA S, WANG J, GE Z, et al. Deep-HOSeq: Deep Higher Order Sequence Fusion for Multimodal Sentiment Analysis[EB/OL]. (2020-10-16)[2021-05-03]. https://arxiv.org/abs/2010.08218.pdf

[20] TACHIBANA H, UENOYAMA K, AIHARA S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, April 15-20, 2018. Washington D.C., USA: IEEE, 2018: 4784-4788.

[21] EYBEN F, WÖLLMER M, SCHULLER B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]//Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, Oct 25-29, 2010. New York, USA: ACM, 2010: 1459-1462.

包广斌（1975-），男，甘肃兰州，博士，副教授，主要研究方向为大数据分析、自然语言处理。

BAO Guangbin, born in 1975, Ph.D. associate professor. His research interests include big data analysis, natural language processing.

李港乐（1997-），女，山东济宁，硕士，主要研究方向为自然语言处理。

LI Gangle, born in 1997, M.S. Her research interests include natural language processing.

王国雄（1997-），男，甘肃陇南，硕士，主要研究方向为自然语言处理。

WANG Guoxiong, born in 1997, M.S. His research interests include natural language processing.