计算机科学
COMPUTER SCIENCE

基于多模态表示学习的情感分析框架

胡新荣, 陈志恒, 刘军平, 彭涛, 叶鹏, 朱强

引用本文

胡新荣, 陈志恒, 刘军平, 彭涛, 叶鹏, 朱强基于多模态表示学习的情感分析框架[J]. 计算机科学, 2022, 49(11A): 210900107-6.

HU Xin-rong, CHEN Zhi-heng, LIU Jun-ping, PENG Tao, YE Peng, ZHU Qiang. Sentiment Analysis Framework Based on Multimodal Representation Learning [J]. Computer Science, 2022, 49(11A): 210900107-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900107

基于多模态表示学习的情感分析框架

胡新荣　陈志恒　刘军平　彭　涛　叶　鹏　朱　强
纺织服装智能化湖北省工程研究中心　武汉 430200
湖北省服装信息化工程技术研究中心　武汉 430200
武汉纺织大学计算机与人工智能学院　武汉 430200
(hxr@wtu.edu.cn)

摘　要　在多模态表示对整体损失的学习过程中,重构损失对模型的依赖性相对较小,导致隐含表示无法有效捕捉它们各自模态的细节。文中提出了一个基于多模态表示学习的多子空间情感分析框架。首先将每个模态投射到模态不变和模态特定两种不同的话语表示中,在模态不变表示中构建主共享子空间以及帮助该子空间减少模态差距的辅助共享子空间,在模态特定表示中构建私有子空间以捕获每个模态独有的特征,将所有子空间中的隐藏向量作为解码函数的输入并重构模态向量,以实现对重构损失的优化。然后,在融合阶段对每个模态表示执行基于 Transformer 的自注意力,使每个表示能从对整体情感取向具有协同作用的其他跨模态表示中获取潜在信息。最后,通过串联生成联合向量并利用全连接层生成任务预测。在两个公开数据集 MOSI 和 MOSEI 上的实验结果表明,该框架在大多数评价指标上都优于基线模型。

关键词:多模态表示;情感分析;Transformer;自注意力;跨模态
中图法分类号　TP391.41

Sentiment Analysis Framework Based on Multimodal Representation Learning

HU Xin-rong,CHEN Zhi-heng,LIU Jun-ping,PENG Tao,YE Peng and ZHU Qiang
Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion,Wuhan Textile University,Wuhan 430200,China
Engineering Research Center of Hubei Province for Clothing Information,Wuhan Textile University,Wuhan 430200,China
School of Computer Science and Artificial Intelligence,Wuhan Textile University,Wuhan 430200,China

Abstract　In the process of learning the overall loss of multimodal representations,the dependence of reconstruction loss on the model is relatively less,resulting in hidden representations that cannot effectively capture the details of their respective modalities. This paper proposes a multi-subspace sentiment analysis framework. Firstly,the framework projects each modality to two distinct utterance representations:modality-invariant and modality-specific. We construct the main shared subspace and the auxiliary shared subspace that helps the main subspace to reduce the modality gap in the modality-invariant representation. Also,construct the private subspaces in the modality-specific representation to capture the characteristic features of each modality. We take the hidden vectors in all subspaces as the input of the decoder function and reconstruct the modal vector to achieve optimization of reconstruction loss. Secondly,in the fusion procedure,we perform a multi-headed self-attention based on Transformer on these representations,so that each cross-modal representation can induce potential information from fellow representations that have a synergistic effect on the overall emotional orientation. Finally,we construct a joint-vector by using concatenation and use fully connected layers to generate task predictions. Experimental results on both MOSI and MOSEI datasets show that the proposed framework outperforms the baselines in most evaluation criteria.

Keywords　Multimodal representation,Sentiment analysis,Transformer,Self-attention,Cross-modality

1　引言

情感分析领域的早期研究工作集中在基于文本的情感分析。然而,在存在大量非语言行为的情况下,说话者所说语句的意义往往会根据非语言行为而动态变化,仅依赖文本进行情感分析不足以准确识别人类表达出的复杂情感[1]。结合声学信息中揭示出单词的音调和语音质量,以及视觉信息中提取出说话者的手势、身体姿势、面部表情等来自其他模态提供的额外情绪信息,有助于消除语句含义中的歧义。

随着 YouTube,Twitter,Facebook 等社交媒体平台的

基金项目:国家自然科学基金(61103085);湖北省高等学校优秀中青年科技创新团队计划项目(T201807);湖北省高校知识产权推进工程项目(GXYS2018009);湖北省教育厅科学研究计划重点项目(D20191708)
This work was supported by the National Natural Science Foundation of China(61103085),Hubei Provincial Outstanding Young and Middle-aged Scientific and Technological Innovation Team Program(T201807),Hubei Province University Intellectual Property Promotion Project(GXYS2018009) and Major Project of Hubei Province Education Department Scientific Research Fund(D20191708).
通信作者:刘军平(jpliu@wtu.edu.cn)

发展，越来越多的用户将这些平台作为日常生活中主要的交流媒介，通过图像、音频和文本等丰富的渠道表达情感，既保证了多模态数据产生的来源，也促使了多模态数据集的创建，例如 CMU-MOSI[2] 和 CMU-MOSEI[3] 等。大多数人工智能研究人员将他们的研究转向多模态情感分析，以利用来自多个来源的各种信息来构建更高效的系统[1]。与传统的单模态情感研究相比，多模态情感分析侧重于对特定模态内动力学和跨模态动力学进行建模[4]。因此，如何实现模态内部的完整表示以及选择最佳的融合方法对不同模态特征进行融合，是研究人员当前面临的挑战。

针对多模态情感分析开发的复杂融合机制无法有效解决相互异质的不同模态之间存在的模态差距，为了更好地利用不同模态之间的互补信息并最大限度地减少模型训练过程中的冗余信息，文献[5]提出了 MISA 多模态框架来为每种模态学习分解子空间，并提供更好的表示作为融合的输入。本文在文献[5]的基础上提出了一个多子空间情感分析框架(Multi-Subspaces Sentiment Analysis，MSSA)，MSSA 首先采用堆叠的双向长短期记忆网络(Long Short Term Memory Network，LSTM)对话语序列进行话语级表示。其次，将固定大小的话语向量投射到模态不变和模态特定两种不同的话语表示中。模态不变表示旨在减小模态之间的差异，一个话语的所有模态表示都投射到具有分布对齐的共享子空间中。模态特定表示通过学习每个模态独有的模态特征，补充模态不变表示中捕获的共同潜在特征的同时，也实现了话语整体的多模态表示。然后，以最小化损失函数为目标，促使模型学习这些子空间表示，损失函数由相似性损失、差异性损失、重构损失和任务预测损失组成。最后，对子空间中的模态向量执行基于 Transformer[6] 的多头自注意力，并以串联方法获得最终的联合向量，再通过全连接层获得任务预测。本文的主要贡献如下：

(1)提出了一个多子空间情感分析框架 MSSA，在模态不变表示中构建辅助共享子空间并将所有子空间中的隐藏向量作为解码函数的输入，以实现重构损失的优化，从而确保隐藏的表示能够更好地捕获它们各自模态的细节，并在融合阶段对所有表示执行多头自注意力。

(2)在 MOSI 和 MOSEI 两个公开数据集上的实验结果表明，MSSA 框架在大多数评价指标上都高于现有基线模型，并通过进一步的消融实验验证了本文框架的有效性。

2 相关工作

多模态情感分析的早期工作主要集中在使用复杂的融合机制学习跨模态动力学[5]。文献[7]提出了一种张量融合网络，可以端到端地对模态间动力学以及模态内动力学进行建模。然而，利用张量的表达能力进行多模态表示，通常会受到维度的指数增长和输入转换为张量的计算复杂度的影响，导致模型面临过度拟合的风险。文献[8]提出了低秩权重张量，在不影响性能的情况下使多模态融合更加高效，在减少参数量的同时也降低了计算复杂度。文献[9]提出了一种融合策略，将整体张量划分为多个局部张量，对特征进行分层融合，相比其他基于张量的方法，该方法可以显著降低计算复杂度。上述的研究仅将每个话语视为一个独立的实体，忽略了相邻话语之间的依赖关系。

最近的一些研究对相邻话语之间的上下文关系展开研究。文献[10]提出了一种基于 LSTM 的模型，该模型使话语能够在同一视频中从其周围环境中捕获上下文信息，从而有助于分类过程。文献[11]使用基于双向循环神经网络的模型提取相邻话语之间的上下文。由于相邻话语在目标话语的情感分类中的重要性存在差异，文献[12]使用上下文感知注意力模块获取上下文信息，并根据相邻话语在当前话语的预测任务中的贡献大小来计算所有相邻话语的注意力权重，以促使网络正确学习话语的局部上下文信息以及视频的全局上下文信息。文献[13]提出了一个基于上下文注意的网络，对话语之间的上下文关系进行建模，并优先考虑重要的上下文信息。

由于现有基于神经网络的方法大部分以隐式和难以理解的方式对多模态交互进行建模，模型可解释性低[1]。文献[14]将量子理论应用于情感分析任务中。文献[15]通过来自量子理论的灵感来表述单一模态内的相互作用和跨模态的相互作用。文献[16]提出了一种受量子干涉启发的多模态决策融合方法和一个受量子测量启发的强弱影响模型，前者用于模拟不同模态之间的决策相关性，后者用于更好地推断说话者之间的社会影响。

文献[5]强调在融合前表征学习的重要性并提出一个多模态情感框架，将模态分解为模态不变和模态特定特征，并将它们融合起来预测情感状态。实验结果证明，通过表示学习函数可以学习到一些理想的特征，有效提高情感预测性能。与上述工作不同的是，本文在情感分析框架中添加辅助共享子空间，为解码函数提供更多的输入，可以帮助重构损失在计算过程中促使隐藏的表示更好地捕获各自模态的细节；并在融合阶段基于跨模态学习方法对所有表示执行多头自注意力，使每个向量都学习到其他的跨模态表示。

3 多子空间情感分析框架及构建方法

3.1 问题定义

本文探讨的问题是利用多模态信号检测视频中的情感。每个视频以话语为单位分割成更小的视频，话语是一种以呼吸或停顿为界限的言语单位[17]。每个话语 U 被视为模型的输入，由文本、视觉和声学 3 个模态对应的低级特征序列组成，分别表示为 $\boldsymbol{U}_l \in \mathbb{R}^{T_l \times d_l}$，$\boldsymbol{U}_v \in \mathbb{R}^{T_v \times d_v}$ 和 $\boldsymbol{U}_a \in \mathbb{R}^{T_a \times d_a}$。$T_m$ 表示话语的长度，d_m 表示模态 m 的特征维数。鉴于这些序列 $\boldsymbol{U}_{m \in \{l,v,a\}}$，通过一组预定义的集合 C 对话语 U 进行情感分类 $y \in \mathbb{R}^C$ 或者作为连续强度变量 $y \in \mathbb{R}$ 进行情感预测。

3.2 MSSA 的框架图

MSSA 框架分为模态表示学习和模态融合两个部分，完整的框架如图 1 所示。

(1)模态表示学习

1)话语级表示

首先，通过一个堆叠的双向 LSTM，将每个模态 $m \in \{l, v, a\}$ 对应的话语序列 $\boldsymbol{U}_m \in \mathbb{R}^{T_m \times d_m}$ 变成一个固定大小的向量 $\boldsymbol{u}_m \in \mathbb{R}^{d_h}$，LSTM 网络的最终状态隐藏表示与完全连接的密集层相结合。该网络的拓扑结构由 LSTM 层、层归一化 Layer-Norm 以及将 ReLU 作为激活函数的全连接层 FC-Layer 组成。

$$\boldsymbol{u}_m = \mathrm{sLSTM}(\boldsymbol{U}_m; \theta_m^{lstm}) \tag{1}$$

2)模态不变表示和模态特定表示

将每一个话语向量 $\boldsymbol{u}_m$ 投射到模态不变和模态特定两个

不同的表示。模态不变表示在具有分布相似性约束的公共子空间中学习共享表示[18]。模态特定表示捕获每个模态的独自特征。

使用编码函数学习给定话语向量 $\boldsymbol{u}_m$ 隐藏的模态不变表示和模态特定表示 $\boldsymbol{h}_m^p \in \mathbb{R}^{d_h}$，模态不变表示构建主共享子空间 $\boldsymbol{h}_m^{c_1} \in \mathbb{R}^{d_h}$ 和辅助共享子空间 $\boldsymbol{h}_m^{c_2} \in \mathbb{R}^{d_h}$。

$$\begin{aligned} h_m^{c_1} &= E_{c_1}(u_m;\theta^{c_1}) \\ h_m^{c_2} &= E_{c_2}(u_m;\theta^{c_2}) \\ h_m^p &= E_p(u_m;\theta_m^p) \end{aligned} \tag{2}$$

然后，使用前馈神经层生成 9 个隐藏向量 $\boldsymbol{h}_{l/v/a}^{c_1/c_2/p}$。$E_{c_1}$ 和 E_{c_2} 为不同模态之间的共享参数 θ^{c_1} 和 θ^{c_2}，E_p 为每个模态分配独立的参数 θ_m^p。编码函数的拓扑结构为将 Sigmoid 作为激活函数的全连接层 FC-Layer，其中包含 128 个神经元。

(2)模态融合

将每个模态投射到各自的表示中，执行基于 Transformer 的自注意力。然后，对 9 个变换后的模态向量进行串联。

Transformer 利用了点积自注意力模块：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V} \tag{3}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别代表查询矩阵、键矩阵和值矩阵。Transformer 计算多个这样的并行注意力，每个多头模块的计算过程如下：

$$head_i = \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^q, \boldsymbol{K}\boldsymbol{W}_i^k, \boldsymbol{V}\boldsymbol{W}_i^u) \tag{4}$$

其中，$\boldsymbol{W}_i^{q/k/v} \in \mathbb{R}^{d_h \times d_h}$ 分别表示对 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 进行变换的矩阵。

1)融合过程

首先，将 9 个模态表示堆叠成一个矩阵 $\boldsymbol{M} = [\boldsymbol{h}_l^{c_1}, \boldsymbol{h}_v^{c_1}, \boldsymbol{h}_a^{c_1}, \boldsymbol{h}_l^{c_2}, \boldsymbol{h}_v^{c_2}, \boldsymbol{h}_a^{c_2}, \boldsymbol{h}_l^p, \boldsymbol{h}_v^p, \boldsymbol{h}_a^p] \in \mathbb{R}^{9 \times d_h}$。然后，执行多头自注意力生成一个新的矩阵 $\bar{\boldsymbol{M}} = [\bar{\boldsymbol{h}}_l^{c_1}, \bar{\boldsymbol{h}}_v^{c_1}, \bar{\boldsymbol{h}}_a^{c_1}, \bar{\boldsymbol{h}}_l^{c_2}, \bar{\boldsymbol{h}}_v^{c_2}, \bar{\boldsymbol{h}}_a^{c_2}, \bar{\boldsymbol{h}}_l^p, \bar{\boldsymbol{h}}_v^p, \bar{\boldsymbol{h}}_a^p] \in \mathbb{R}^{9 \times d_h}$。设置 $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{M} \in \mathbb{R}^{6 \times d_h}$，每一个 $head_i$ 都通过式(4)计算得到，符号⊕代表串联，$\theta^{att} = \{W^q, W^k, W^v, W^o\}$。$\bar{\boldsymbol{M}}$ 的表达式如式(5)所示：

$$\bar{\boldsymbol{M}} = \mathrm{MultiHead}(\boldsymbol{M};\theta^{att}) = (head_1 \oplus \cdots \oplus head_n)W^o \tag{5}$$

2)任务预测

通过串联构造联合向量 $\boldsymbol{h}^{\mathrm{out}} = [\bar{\boldsymbol{h}}_l^{c_1} \oplus \cdots \oplus \bar{\boldsymbol{h}}_a^p] \in \mathbb{R}^{9d_h}$，其中符号⊕代表串联运算。然后由全连接层 $\hat{\boldsymbol{y}} = G(\boldsymbol{h}^{\mathrm{out}};\theta^{\mathrm{out}})$ 生成任务预测。

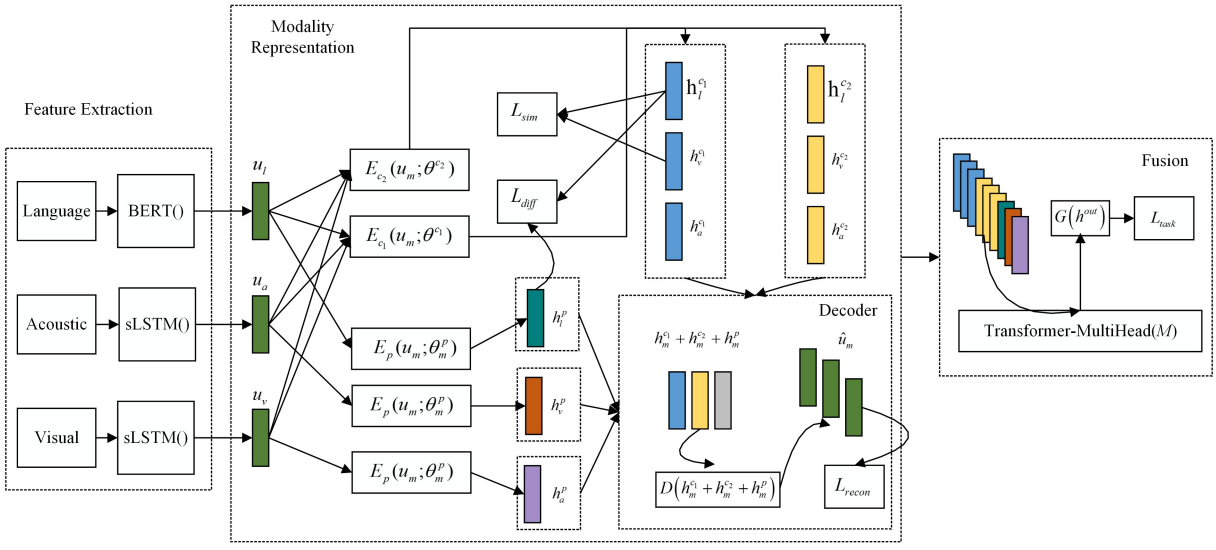

图 1　MSSA 整体框架图

Fig. 1　Overall framework of MSSA

3.3　损失函数

模型的整体学习通过最小化式(6)来执行。

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \alpha\mathcal{L}_{\mathrm{sim}} + \beta\mathcal{L}_{\mathrm{diff}} + \gamma\mathcal{L}_{\mathrm{recon}} \tag{6}$$

其中，α,β,γ 代表每个单独损失对整体损失 $\mathcal{L}$ 的贡献的权重因子。

3.3.1　相似性损失

使用距离度量方法 CMD(Central Moment Discrepancy)[19]最小化相似性损失，以减小两个模态对应共享表示之间的差异，CMD 的定义如下：

$$CMD_K(X,Y) = \frac{1}{|b-a|}\|\boldsymbol{E}(X)-\boldsymbol{E}(Y)\|_2 + \sum_{k=2}^{K}\frac{1}{|b-a|^k}\|C_k(X)-C_k(Y)\|_2 \tag{7}$$

其中，$\boldsymbol{E}(X) = \frac{1}{|X|}\sum_{x \in X} x$ 是样本 X 的经验期望向量，$C_k(X) = \boldsymbol{E}((x-\boldsymbol{E}(X))^k)$ 是 X 坐标的所有 k^{th} 阶样本中心矩的向量。

本文中，计算模态不变表示之间的 CMD 损失表达式如下：

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{3}\sum_{(m_1,m_2)\in\{(l,a),(l,v),(a,v)\}} CMD_K(\boldsymbol{h}_{m_1}^{c_1}, \boldsymbol{h}_{m_2}^{c_1}) \tag{8}$$

3.3.2　差异损失

差异损失是为了确保模态不变表示和特定表示捕获输入的全面性。在一轮训练中，将矩阵 $\boldsymbol{H}_m^{c_1}$ 和 $\boldsymbol{H}_m^p$ 变换为具有零均值和单位 $l2$ 范数的矩阵，其行表示每个话语对应的模态 m 的隐藏向量 $\boldsymbol{h}_m^{c_1}$ 和 $\boldsymbol{h}_m^p$。计算模态向量之间的正交约束如下：

$$\|\boldsymbol{H}_m^{c_1\mathrm{T}}\boldsymbol{H}_m^p\|_F^2 \tag{9}$$

其中，$\|\cdot\|_F^2$ 是 Frobenius 范数的平方。同一模态的不变向量和特定向量之间以及不同模态的特定向量之间都添加正交性约束，总体差异损失的计算如下：

$$\mathcal{L}_{\mathrm{diff}} = \sum_{m\in\{l,v,a\}}\|\boldsymbol{H}_m^{c_1\mathrm{T}}\boldsymbol{H}_m^p\|_F^2 + \sum_{(m_1,m_2)\in\{(l,a),(l,v),(a,v)\}}\|\boldsymbol{H}_{m_1}^{p\mathrm{T}}\boldsymbol{H}_{m_2}^p\|_F^2 \tag{10}$$

3.3.3　重构损失

重构损失是为了确保隐藏的表示能够捕获它们各自模态的细节。使用解码器函数 $\hat{\boldsymbol{u}}_m = D(\boldsymbol{h}_m^{c_1} + \boldsymbol{h}_m^{c_2} + \boldsymbol{h}_m^p;\theta^d)$ 重建模态

向量 $\boldsymbol{u}_m$，并计算 $\boldsymbol{u}_m$ 和 $\hat{\boldsymbol{u}}_m$ 之间的均方误差损失：

$$\mathcal{L}_{\text{recon}}=\frac{1}{3}\left(\sum_{m\in\{l,v,a\}}\frac{\|\boldsymbol{u}_m-\hat{\boldsymbol{u}}_m\|_2^2}{d_h}\right) \tag{11}$$

其中，$\|\cdot\|_2^2$ 是 $L2$-范数的平方。

3.3.4 任务损失

特定任务损失是为了估计训练中预测的质量。分类任务中使用标准交叉熵损失；回归任务中使用均方误差损失。对一轮训练中的话语 N_b 的计算式如下，其中式(12)用于分类任务，式(13)用于回归任务：

$$\mathcal{L}_{\text{task}}=-\frac{1}{N_b}\sum_{i=0}^{N_b}y_i\cdot\log\hat{y}_i \tag{12}$$

$$\mathcal{L}_{\text{task}}=\frac{1}{N_b}\sum_{i=0}^{N_b}\|y_i-\hat{y}_i\|_2^2 \tag{13}$$

4 实验

4.1 数据集

MSSA 框架在两个由卡梅隆大学所提供的基准数据集 MOSI 和 MOSEI 上进行实验和评估。数据集中为每个话语提供单词对齐的多模态信号(语言、视觉和听觉)。

MOSI 数据集中收集的视频来自于 YouTube，从 89 位不同的演讲者中选择了 93 个视频，其中，男性演讲者 47 位，女性演讲者 41 位。MOSEI 是一个规模更大的数据集，由 3228 个视频组成，共包含由 1000 多名在线 YouTube 演讲者提供的 22777 条话语，讨论了 250 个不同的主题。数据集的训练、验证和测试的分割情况如表 1 所列。

表 1 数据集的分割情况

Table 1 Datasets segmentation in experiment

Datasets	train	dev	test
MOSI	1283	229	686
MOSEI	16315	1871	4654

数据集中的话语情绪强度被手工定义为强烈积极(+3)，积极(+2)，弱积极(+1)，中性(0)，弱消极(-1)，消极(-2)和强烈消极(-3)。

4.2 评价指标

我们在两个数据集的基础上执行分类和回归任务。分类任务的评价指标包括 seven-class accuracy(Acc-7)，binary accuracy(Acc-2)和 F-Score。其中，binary accuracy 指标使用两种方法，第一种是负/非负，非负标签代表情感分数大于等于 0[20]；第二种是负/正，分别代表小于 0 和大于 0 的情绪分数[21]。实验结果的表格中，-/-的左侧代表负/非负的结果，右侧代表负/正的结果。回归任务中将 Mean Absolute Error(MAE)和 Pearson correlation(Corr)作为评价指标，MAE 指标越低越好。

4.3 特征提取

4.3.1 文本特征

将 BERT-base-uncased 预训练模型[22]作为文本特征提取器，该模型由 12 个堆叠的 Transformer 层组成，我们从最终的 768 维隐藏状态中选择话语向量 u_l 作为标记的平均表示，计算式如下：

$$\boldsymbol{u}_l=\text{BERT}(\boldsymbol{U}_l;\theta^{\text{bert}}) \tag{14}$$

4.3.2 视觉特征

使用 facet 工具提取面部表情特征，其中包括面部动作单元和基于面部动作编码系统(FACS)[23]的面部姿态。MOSI 和 MOSEI 数据集上视觉特征的输入维度 d_v 分别是 47 和 35。

4.3.3 声学特征

通过 COVAREP[24]声学分析框架，提取各种低级统计音频函数，包括 12 Mel-frequency ceptral coefficients，pitch，Voiced/Unvoiced segmenting features (VUV)[25]，glottal source parameters[26]等。MOSI 和 MOSEI 数据集上声学特征的输入维度 d_a 都是 74。

4.4 基线模型

为了评估本文模型在分类和回归任务上的性能，与下述模型进行全面的比较。使用的基线模型如下：1)RMFN[27]，将融合问题分解为多个阶段，通过多阶段融合方法对跨模态交互进行建模；2)RAVEN[28]，利用基于注意力的非语言信号模型捕捉给定非语言上下文的词，表示空间中有意义的动态变化；3)MCTN[29]，通过从源到目标模态的循环转换来学习联合表示；4)MFM[30]，通过模态分解模型将多模态表示分解为多模态判别因子和模态特定生成因子；5)MulT[21]，通过跨模态注意机制直接关注其他模态中的低级特征来融合多模态信息；6)TFN[7]，将多模态情感分析问题作为模态内和模态间动态建模，通过张量融合网络端到端地学习这两种动态。7)LMF[8]，使用低秩张量执行多模态融合以提高效率；8)CIA[12]，通过自动编码器学习模态间的交互关系，并采用上下文感知注意模块学习相邻话语之间的对应关系；9)ICCN[31]，通过深度典型相关分析学习所有 3 种模式之间的相关性；10)MISA[5]，通过学习有效的模态表示来帮助融合过程，我们重新做了实验，与本文提出的 MSSA 框架进行了公平的对比。

4.5 训练细节

为了两个数据集分别设置合适的超参数。MOSI 数据集中，batch size 为 64，dropout 为 0.1，激活函数使用 ReLU；MOSEI 数据集中，batch size 为 16，dropout 为 0.5，激活函数使用 LeakyReLU。α,β,γ 的值分别为 1.0，0.3，1.0，学习率为 1×10^{-4}。本文实验代码已公布到相关网站[1)]。

5 实验结果与分析

表 2 和表 3 列出了不同模型在 MOSI 和 MOSEI 数据集上的实验结果。总体来说，本文提出的 MSSA 框架在两个数据集上的表现优于其他基线模型。表 2 中，SOTA1 表示本文提出的 MSSA 框架与所有基线模型中评价指标最优的对比结果，我们发现除了使用负/正分类方法的 Acc-2 和 F-Score 两个指标降低了 0.4 外，其余的指标均有提高。SOTA2 以及表 3 中的 SOTA 表示 MSSA 框架与基线模型 MISA 的对比结果。其中各项指标均超过了基线模型 MISA，这表明本文添加的辅助共享子空间可以有效提高情感分析任务的性能。此外，为了分析不同距离度量方法对模型性能的影响，我们还使用另一种距离度量方法 MMD，代替 CMD 作为相似性损

1) https://github.com/qqhh0830/MSSA

失函数。实验结果 MSSA-mmd 表明，距离度量方法的不同会对模型产生很大影响，使用 MMD 方法会导致模型性能下降。这是由于 MMD 公式相对复杂并且耗费了昂贵的距离计算。

表 2　MOSI：MSSA 框架性能与基线模型对比结果

Table 2　Performance comparison of MSSA framework and baselines in MOSI dataset

Models	Acc-2	F-Score	Acc-7	MAE	Corr
RMFN	78.4/	78.0/	38.3	0.922	0.681
RAVEN	78.0/	76.6/	33.2	0.915	0.691
MCTN	79.3/	79.1/	35.6	0.909	0.676
CIA	79.8/	79.5/	38.9	0.914	0.689
MulT	/83.0	/82.8	40.0	0.871	0.698
TFN	/80.8	/80.7	34.9	0.901	0.698
LMF	/82.5	/82.4	33.2	0.917	0.695
MFM	/81.7	/81.6	36.2	0.877	0.706
ICCN	/83.0	/83.0	38.9	0.860	0.710
MISA	77.8/80.5	77.8/80.6	42.7	0.799	0.737
MMD	77.6/79.6	77.5/79.6	39.9	0.872	0.705
MSSA	80.2/82.6	80.0/82.6	43.6	0.792	0.745
SOTA1	0.4↑/0.4↓	0.5↑/0.4↓	0.9↑	0.007↑	0.008↑
SOTA2	2.4↑/2.1↑	2.2↑/2.0↑	0.9↑	0.007↑	0.008↑

表 3　MOSEI：MSSA 框架性能与基线模型对比结果

Table 3　Performance comparison of MSSA framework and baseline model in MOSEI dataset

Models	Acc-2	F-Score	Acc-7	MAE	Corr
RAVEN	79.1/	79.5/	50.0	0.614	0.662
MCTN	79.8/	80.6/	49.6	0.609	0.670
CIA	80.4/	78.2/	50.1	0.680	0.590
MulT	/82.5	/82.3	51.8	0.580	0.703
TFN	/82.5	/82.1	50.2	0.593	0.700
LMF	/82.0	/82.1	48.0	0.623	0.677
MFM	/84.4	/84.3	51.3	0.568	0.717
ICCN	/84.2	/84.2	51.6	0.565	0.713
MISA	81.6/84.7	82.1/84.7	52.2	0.551	0.754
MMD	81.9/84.8	82.4/84.9	51.3	0.556	0.749
MSSA	83.3/85.6	83.6/85.5	53.2	0.542	0.759
SOTA	1.7↑/0.9↑	1.5↑/0.8↑	1.0↑	0.009↑	0.005↑

在消融实验中，为了验证 3 个损失函数的重要性，通过每次消除一个损失来重新训练两个数据集对应的最佳模型。表 4 中第 1—3 行的实验结果表明，当使用 3 个损失函数时，模型获得了最佳的性能。

为了分析子空间的作用，提出了 4 种模型的变体进行实验：1）删除模态不变子空间，只学习模态特定表示，然后用于融合（MSSA-MS）；2）删除模态特定子空间，只学习模态不变表示，然后用于融合（MSSA-MI）；另外两个变体在表征学习阶段与 MSSA 相同；3）在融合过程中，仅使用模态特定表示进行融合和预测（MSSA-msFusion）；4）仅使用模态不变表示进行融合和预测（MSSA-miFusion）。表 4 中第 4、5 行的实验结果表明，在该多模态表示学习框架中，学习模态特定表示比模态不变表示更有效，在回归任务中性能表现得更好。这是由于仅学习模态不变表示的过程中，话语的不同模态之间并不都存在有效的可共享信息。表 4 中第 6、7 行的实验结果表明，在融合阶段仅使用模态特定特征进行融合，模型预测性能高于仅使用模态不变特征进行融合，这是由于模态不变特征中包含了更多的情感信息。为了进一步验证辅助共享子空间的贡献，对融合阶段的矩阵中的辅助共享子空间的模态表示进行删除，$M=\lceil h_l^{c_l}, h_v^{c_l}, h_a^{c_l}, h_l^p, h_v^p, h_a^p \rceil \in \mathbb{R}^{6\times d_h}$，即 MSSA-X。表 4 中第 8 行的实验结果表明，相比 MSSA，两项指标都发生了下降，在融合过程中加入辅助子空间中的模态表示对预测性能产生了影响。

通过以上消融实验结果表明，学习模态不变和模态特定两种不同的话语表示，并在融合阶段对所有模态特征进行融合，模型可以达到最佳的性能，这证实了本文提出的 MSSA 框架的有效性。

此外，在 MOSEI 数据集上的各项指标都优于 MOSI 数据集，原因是 MSSA 框架对多模态表示的学习更依赖大量的数据集来支撑。

表 4　MSSA 框架消融实验结果

Table 4　Ablation experiment results of MSSA framework

Number	Models	MOSI		MOSEI	
		MAE	Corr	MAE	Corr
1	MSSA-α	0.923	0.710	0.547	0.758
2	MSSA-β	0.828	0.738	0.540	0.762
3	MSSA-γ	0.874	0.708	0.551	0.756
4	MSSA-MS	0.821	0.743	0.551	0.757
5	MSSA-MI	0.836	0.723	0.554	0.758
6	MSSA-msFusion	0.857	0.714	0.555	0.755
7	MSSA-miFusion	0.849	0.738	0.548	0.760
8	MSSA-X	0.869	0.723	0.551	0.753

结束语　本文提出了一种基于多模态表示学习的多子空间情感分析框架 MSSA，通过在模态不变表示中添加辅助共享子空间，来弥补独立的共享子空间无法有效促使跨模态表示学习模态之间的共性，并通过优化重构损失来帮助隐藏表示捕获各自模态的细节，同时提高了模型对重构损失的依赖性。在公开数据集上的对比实验和消融实验验证了本文提出的框架以及辅助共享子空间的有效性。在融合阶段直接通过串联来构造联合向量，无法充分利用表示学习函数学习到的理想特征，在未来会尝试使用不同的融合方法对融合阶段进行改进。

参 考 文 献

[1] ABDU S A，YOUSEF A H，SALEM A. Multimodal video sentiment analysis using deep learning approaches，a survey[J]. Information Fusion，2021，76(2021)：204-226.

[2] ZADEH A，ZELLERS R，PINCUS E，et al. Mul-timodal sentiment intensity analysis in videos：facial gestures and verbal messages[J]. IEEE Intelligent Systems，2016，31(6)：82-88.

[3] ZADEH A，LIANG P P，VANBRIESEN J，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. 2018：2236-2246.

[4] RAJAGOPALAN S S，MORENCY L P，BALTRUSAITIS T，et al. Extending long short-term memory for multi-view structured learning [C] // European Conference on Computer Vision. 2016：338-353.

[5] HAZARIKA D，ZIMMERMANN R，PORIA S. MISA：modality-invariant and -specific representations for multimodal sentiment analysis [C] // ACM International Conference on Multimedia. 2020：1122-1131.

[6] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is

all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
[7] 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. 2017: 1103-1114.
[8] 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. 2018: 2247-2256.
[9] MAI S, HU H, XING S, et al. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 481-492.
[10] 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. 2017: 873-883.
[11] 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. 2018: 3454-3466.
[12] CHAUHAN D S, AKHTAR M S, EKBAL A, et al. Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis [C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 5647-5657.
[13] PORIA S, CAMBRIA E, HAZARIKA D, et al. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis [C] // IEEE International Conference on Data Mining. 2017: 1033-1038.
[14] ZHANG Y Z, SONG D W, ZHANG P. A Quantum-Inspired Multimodal Sentiment Analysis Framework [J]. Theoretical Computer Science, 2018, 752(2018): 21-40.
[15] LI Q C, GKOUMAS D, LIOMA C, et al. Quantum-inspired Multimodal Fusion for Video Sentiment Analysis [J]. Information Fusion, 2021, 65(2021): 58-71.
[16] ZHANG Y, SONG D, LI X, et al. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis [J]. Information Fusion, 2020, 62(2020): 14-31.
[17] OLSON D. From utterance to text: The bias of language in speech and writing [J]. Harvard Educational Review, 1977, 47(3): 257-281.
[18] GUO W Z, WANG J W, WANG S P. Deep Multimodal Representation Learning: A Survey [J]. IEEE Access, 2019, 7(2019): 63373-63394.
[19] ZELLINGER W, LUGHOFER E, SAMINGER-PLATZ S, et al. Central moment discrepancy (cmd) for domain-invariant representation learning[J]. arXiv: 1702.08811, 2017.
[20] ZADEH A, LIANG P P, PORIS S, et al. Multi-attention Recurrent Network for Human Communication Comprehension [C] // AAAI Conference on Artificial Intelligence. 2018: 5642-5649.
[21] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences [C] // Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019: 6558-6569.
[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J] arXiv: 1810.04805, 2018.
[23] EKMAN P, ROSENBERG E L. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS) [M] // USA: Oxford University Press.
[24] DEGOTTEX G, KANE J, DRUGMAN T, et al. Cvarep-a collaborative voice analysis repository for speech technologies [C] // 2014 IEEE International Conference on Acoustics. Speech and Signal Processing, 2014: 960-964.
[25] DRUGMAN T, ALWAN A. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics [J]. arXiv: 2001.00459, 2019.
[26] DRUGMAN T, THOMAS M, GUDNASON J, et al. Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(3): 994-1006.
[27] LIANG P P, LIU Z Y, ZADEH A, et al. Multimodal Language Analysis with Recurrent Multistage Fusion [J]. arXiv: 1808.03920, 2018.
[28] WANG Y S, SHEN Y, LIU Z, et al. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7216-7223.
[29] 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. 2019: 6892-6899.
[30] TSAI Y H H, LIANG P P, ZADEH A. Learning factorized multimodal representations[J]. arXiv: 1806.06176, 2018.
[31] SUN Z K, SARMA P K, SETHARES W A, et al. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 8992-8999.

HU Xin-rong, born in 1973, Ph.D, professor, postgraduate supervisor, is a member of China Computer Federation. Her main research interests include image processing and pattern recognition, virtual reality and natural language processing.

LIU Jun-ping, born in 1979, Ph.D, professor, postgraduate supervisor, is a member of China Computer Federation. His main research interests include industrial big data and artificial intelligence.