计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330722
61(5): 1310−1324, 2024

基于样本内外协同表示和自适应融合的多模态学习方法

黄学坚[1,2] 马廷淮[2] 王根生[3]
[1](江西财经大学虚拟现实(VR)现代产业学院 南昌 330013)
[2](南京信息工程大学计算机学院 南京 210044)
[3](江西财经大学信息管理学院 南昌 330013)
(huangxuejian@jxufe.edu.cn)

Multimodal Learning Method Based on Intra- and Inter-Sample Cooperative Representation and Adaptive Fusion

Huang Xuejian[1,2], Ma Tinghuai[2], and Wang Gensheng[3]
[1](*College of VR Modern Industry*, *Jiangxi University of Finance and Economics*, *Nanchang* 330013)
[2](*College of Computer*, *Nanjing University of Information Science and Technology*, *Nanjing* 210044)
[3](*College of Information Management*, *Jiangxi University of Finance and Economics*, *Nanchang* 330013)

Abstract Multimodal machine learning represents a novel paradigm in artificial intelligence, leveraging various modalities and intelligent processing algorithms to achieve enhanced performance. Multimodal representation and fusion are two pivotal tasks in multimodal machine learning. Currently, most multimodal representation methods pay little attention to inter-sample collaboration, leading to a lack of robustness in feature representation. Additionally, most multimodal feature fusion methods exhibit sensitivity to noisy data. Therefore, in the realm of multimodal representation, an approach based on both intra-sample and inter-sample multimodal collaboration is proposed to facilitate a comprehensive understanding of interactions within and between modalities, ultimately enhancing the robustness of feature representation. Firstly, text, speech, and visual features are individually extracted based on pre-trained models such as BERT, Wav2vec 2.0, and Faster R-CNN. Subsequently, considering the complementarity and consistency of multimodal data, two categories of encoders, namely modality-specific and modality-shared, are constructed to learn both modality-specific and shared feature representations. Furthermore, intra-sample collaboration loss functions are formulated using central moment differences and orthogonality, while inter-sample collaboration loss functions are established using contrastive learning. Lastly, a representation learning function is designed based on intra-sample collaboration, inter-sample collaboration, and sample reconstruction errors. Regarding multimodal fusion, an adaptive multimodal feature fusion method is designed, accounting for the possibility that each modality may exhibit varying types of effects and levels of noise at different times, using attention mechanisms and gated neural networks. Experimental results on the multimodal intent recognition dataset MIntRec and emotion datasets CMU-MOSI and CMU-MOSEI demonstrate that this multimodal learning approach outperforms baseline methods across multiple evaluation metrics.

Key words multimodal representation; multimodal fusion; multimodal learning; collaborative representation; adaptive fusion

收稿日期: 2023-09-06; **修回日期**: 2024-02-20
基金项目: 国家自然科学基金项目(62372243, 72061015, 62102187)
This work was supported by the National Natural Science Foundation of China (62372243, 72061015, 62102187).
通信作者: 马廷淮(thma@nuist.edu.cn)

摘　要　多模态机器学习是一种新的人工智能范式，结合各种模态和智能处理算法以实现更高的性能. 多模态表示和多模态融合是多模态机器学习的2个关键任务. 目前，多模态表示方法很少考虑样本间的协同，导致特征表示缺乏鲁棒性，大部分多模态特征融合方法对噪声数据敏感. 因此，在多模态表示方面，为了充分学习模态内和模态间的交互，提升特征表示的鲁棒性，提出一种基于样本内和样本间多模态协同的表示方法. 首先，分别基于预训练的BERT, Wav2vec 2.0, Faster R-CNN提取文本特征、语音特征和视觉特征；其次，针对多模态数据的互补性和一致性，构建模态特定和模态共用2类编码器，分别学习模态特有和共享2种特征表示；然后，利用中心矩差异和正交性构建样本内协同损失函数，采用对比学习构建样本间协同损失函数；最后，基于样本内协同误差、样本间协同误差和样本重构误差设计表示学习函数. 在多模态融合方面，针对每种模态可能在不同时刻表现出不同作用类型和不同级别的噪声，设计一种基于注意力机制和门控神经网络的自适应的多模态特征融合方法. 在多模态意图识别数据集MIntRec和情感数据集CMU-MOSI, CMU-MOSEI上的实验结果表明，该多模态学习方法在多个评价指标上优于基线方法.

关键词　多模态表示；多模态融合；多模态学习；协同表示；自适应融合

中图法分类号　TP391

多模态机器学习旨在建立能够处理和关联来自多种模式信息的模型，近年来成为研究的热点. 多模态表示和多模态融合是多模态机器学习的2个关键任务[1]. 由于模态间的异构性，多模态表示学习一直是个难点问题. 目前，基于神经网络的联合表示学习模型把所有的模态数据映射到统一的特征空间，得到联合特征表示，容易实现端到端的学习，但需要大量的标注数据[2]. 在一些应用领域，多模态数据具有共享和特有的特征，例如在多模态情感识别任务中，说话人的动作、语音和语言具有共同的动机和目标，同时它们又分别具有特有的情感、语气和语义. 为了有效学习不同模态的共享特征和特有特征，Hazarika等人[3]提出了一种多模态协同表示模型MISA，将每个模态映射到2个不同的子空间中，分别学习共享特征和特有特征，但该模型只考虑了单个样本内的多模态协同，没有考虑样本间的多模态协同，导致不同类别样本的特征空间具有一定程度的重合，特征表示缺乏鲁棒性.

多模态融合根据融合阶段的不同，可以分为早期融合、晚期融合和混合融合[4]. 早期融合是特征层的融合，在融合后的特征上训练分类器；晚期融合是决策层的融合，每个模态数据单独训练一个分类器，然后根据投票、加权和学习等方式对分类器的结果进行融合；混合融合联合了早期融合和晚期融合2种方式，试图同时利用2种融合方式的优点. 晚期融合允许不同的模态采用不用的预测模型，使得模型具有灵活性，但忽视了不同模态特征的交互. 早期融合使用单一模型进行训练，实现了不同模态特征的交互. 研究表明[5-6]，在多模态语言分析任务中，文本特征占据了主要地位，语音和视频常为辅助特征，在某些情况下语音和视频可能包含噪声，对结果的判断起到干扰作用. 目前大部分多模态融合方法，把所有的模态特征同等对待，导致对噪声数据敏感.

因此，针对多模态协同表示没有考虑样本间的协同和多模态特征融合对噪声数据敏感的问题，本文提出一种基于样本内外协同表示和自适应融合的多模态学习方法. 在多模态表示方面，构建模态特定和模态共用的2类编码器分别学习文本、视频和语音的特有特征和共享特征的表示，通过样本重构误差、样本内协同误差和样本间协同误差设计表示学习损失函数. 在多模态特征融合方面，设计一种基于注意力机制和门控神经网络的自适应的融合方法，利用注意力机制学习模态间的依赖关系，通过门控神经网络得出融合权重. 在多模态意图识别数据集MIntRec和多模态情感数据集CMU-MOSI, CMU-MOSEI上的实验结果表明，本文提出的多模态学习方法在多个指标上优于基线方法，证明了该方法的有效性.

本文的主要贡献包括3个方面：

1）提出了一种基于样本内和样本间多模态协同的表示方法，充分学习模态内和模态间的交互，提升多模态特征表示的鲁棒性.

2）设计了一种基于注意力机制和门控神经网络的自适应的多模态特征融合方法，降低噪声数据对多模态融合过程的干扰.

3）在多模态意图识别数据集和情感数据集上对本文提出的方法进行了大量的实验分析，本文方法在多个指标上优于基线方法.

1 相关工作

1.1 多模态表示

特征表示一直是机器学习关注的重要问题. 随着深度学习的发展，单模态的特征表示学习取得了很多进展，但由于数据的异构性，多模态表示学习一直是个难点问题[7]. 目前，多模态表示学习主要分为联合表示(joint representations)和协同表示(coordinated representations). 联合表示通过神经网络将各模态数据映射到同一个特征空间中，得到统一的特征表示，使得多模态表示学习和多模态融合之间没有明显的界限. 例如，Pham 等人[8]利用机器翻译的思想，通过 Seq2Seq 模型实现不同模态之间的来回转换，把 Seq2Seq 中间隐含层的输出作为多模态的联合表示；Wang 等人[9]提出一种通过门控模态混合网络实现文本和非文本特征联合表示的方法. 协同表示分别映射每种模态的数据到各自的特征空间，但要保证每种模态的特征空间之间存在一定的约束. 例如，Mai 等人[10]提出一种基于混合对比学习的多模态协同表示方法，首先通过 Transformer 提取语音和视觉特征，通过 BERT 提取文本特征，然后通过模态内对比学习、模态间对比学习和半对比学习对语音特征、视觉特征和文本特征的相似性进行约束；Hazarika 等人[3]提出一种多模态协同表示方法 MISA，将每种模态投射到 2 个不同的子空间，第 1 个子空间是模态不变的，通过相似性进行约束，第 2 个子空间是模态特有的，通过正交结构进行约束；Huang 等人[11]在 MISA 方法的基础上，通过中心矩差异对模态的特征空间进行约束.MISA 是一种有效的多模态协同表示方法，能够很好地学习不同模态的共享特征和特有特征. 然而，MISA 仅考虑了样本内的多模态协同约束，未考虑样本间的多模态协同，导致特征表示缺乏鲁棒性，从而影响模型的泛化能力. 因此，在 MISA 的基础上，本文提出一种基于样本内和样本间多模态协同的表示方法，充分学习模态内和模态间的交互，提升多模态特征表示的鲁棒性.

1.2 多模态融合

多模态融合关注于如何将多模态数据以一定的架构和方法进行融合，共同贡献于解决目标任务[12-13]，多模态融合主要分为模型无关和基于模型 2 类方法[14]. 模型无关的方法主要分为特征层融合和决策层融合. 特征层融合实现了不同模态间的底层交互，常用的融合方式有拼接、相加和基于张量的方法[15-16]；决策层融合可以视为考虑不同模态置信度的集成学习，其优点是能够很好地适应模态缺失的问题[17]，但缺乏多模态数据的底层交互，常用的融合机制有加权、投票和学习等方式. 基于模型的融合方法主要有基于内核的方法、基于概率图模型的方法和基于神经网络的方法[18-20]. 目前，基于神经网络的多模态融合方法已经成为主流[21]，例如：Liang 等人[22]提出一种循环多级融合网络 RMFN，将融合问题分解为多个阶段，每个阶段专注于多模态数据的一个子集；Tsai 等人[23]提出一种多模态 Transformer 架构，通过跨模态注意力机制融合多模态信息；Mou 等人[24]提出一种基于注意力的卷积神经网络(convolutional neural networks, CNN)和长短期记忆网络(long short-term memory, LSTM)联合的多模态融合方法；Rahman 等人[25]为了在大规模预训练语言模型中融合其他模态信息，在 BERT 和 XLNet 网络中设计了一个多模态适应门，允许 BERT 和 XLNet 在微调期间接受多模态非语言数据. 通过研究发现，目前大部分多模态融合方法没有区分模态间可能存在的主次关系，并且没有考虑数据中可能存在的噪声，导致模型对噪声数据敏感. 因此，鉴于每种模态在不同时刻可能呈现不同作用类型和噪声级别的特性，本文设计一种基于注意力机制和门控神经网络的融合方法，以实现对多模态特征的自适应融合.

2 方法构建

本文方法主要面向于文本特征为主、语音和视觉特征为辅的多模态自然语言理解任务，例如多模态情感分类和多模态意图识别. 给定一个数据集 $D=\{s_1,s_2,\cdots,s_n\}$，其中包含 n 个样本. 每个样本 s_i 都包含一段视频 v、语音 a、文本 t 和标签 y. 我们的任务是学习一个模型 $f(t,v,a)\rightarrow y$，输入样本 s_i 的文本信息 t、视频信息 v 和语音信息 a，正确输出样本 s_i 的标签信息 y. 本文提出的基于样本内外协同表示和自适应融合的多模态学习方法 CoAdMu，其架构如图 1 所示，主要包括初始特征提取、多模态表示、多模态融合和结果预测 4 个部分.

2.1 初始特征提取

2.1.1 文本初始特征提取

预训练语言模型能够很好地提取文本语义特征，已经成为自然语言处理任务的标配模块. 预训练语言模型 BERT 基于 Transformer 的双向 Encoder 结构，采用 Self-attention 提高了模型的学习能力和并行计

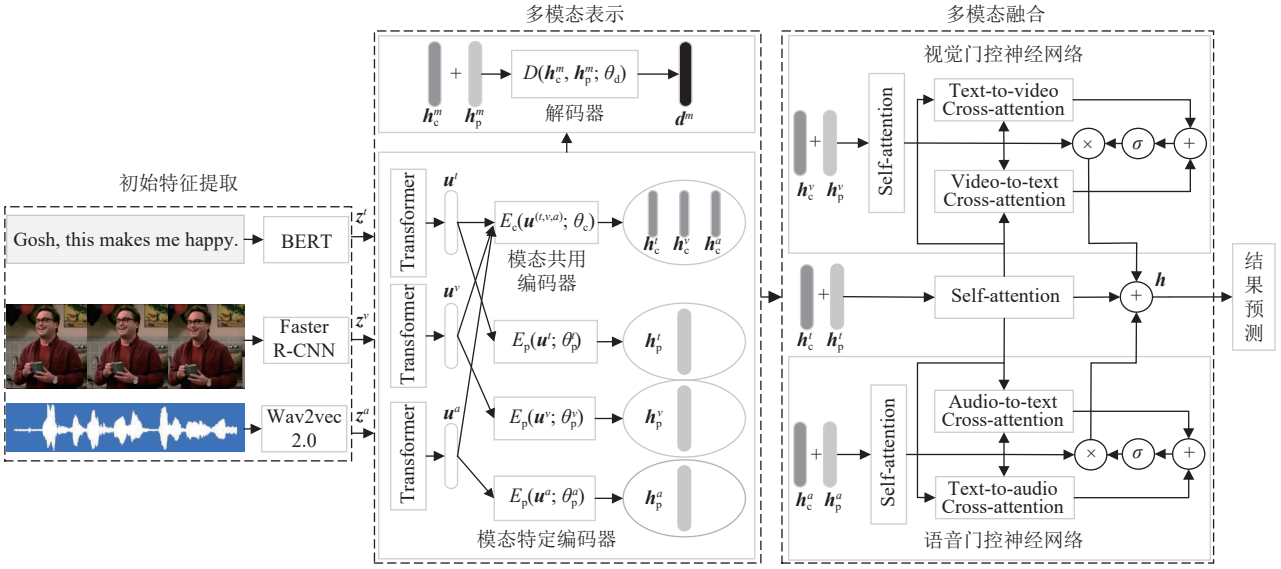

Fig. 1 Architecture of CoAdMu

图 1 CoAdMu 的架构

算效率. 为了获取句子级别的语义特征, BERT 联合了 Masked LM(masked language model)和 NSP(next sentence prediction)这 2 类任务进行训练.BERT 相比于 Word2vec, 一方面考虑了上下文语境, 解决了一词多义的问题; 另一方面, 通过分层学习得到不同层次的语义特征, 为下游任务提供了丰富的特征选择. 基于 BERT 预训练模型, 下游任务可以进行微调, 在少量训练样本的情况下就能实现不错的分类效果. 所以, 本文利用 BERT 最后隐藏层的输出 $\boldsymbol{z}^t \in \mathbb{R}^{l_t \times F_t}$ 作为文本初始特征表示, l_t 为文本序列长度, F_t 为特征维度.

2.1.2 语音初始特征提取

近年来, 受大规模预训练语言模型在自然语言理解任务上大获成功的影响, 语音预训练模型成为研究的热点, 出现了许多经典的模型. 例如 Wav2vec 2.0 [26], HuBERT [27] 和 WavLM[28] 等, 通过在上万小时的无标注语音数据上进行自监督学习, 显著提升了自动语音识别(automatic speech recognition, ASR)、语音合成(text-to-speech, TTS)和语音转换(voice conversation, VC)等下游任务的性能.Wav2vec 2.0 是 Meta 在 2020 年发布的无监督语音预训练模型, 核心思想是通过向量量化(vector quantization, VQ)构建自监督训练目标, 对输入做大量掩码后利用对比学习损失函数进行训练, 得到的表征可以代替传统的声学特征. 所以, 本文利用预训练的 Wav2vec 2.0 模型提取语音初始特征, 把模型最后隐藏层的输出 $\boldsymbol{z}^a \in \mathbb{R}^{l_a \times F_a}$ 作为语音初始特征表示, l_a 为语音序列长度, F_a 为特征维度.

2.1.3 视频初始特征提取

在视频画面中, 关键信息是说话人(识别对象)的表情和动作, 如果直接从整个画面中抽取特征, 可能会因为背景噪声影响效果. 所以本文借鉴文献 [29] 的思路, 对说话人进行检测. 具体流程如图 2 所示.

Fig. 2 说话人检测

图 2 Speaker detection

首先利用场景检测工具 scenedetect① 区分不同的视觉场景, 从而得到关键帧; 然后, 利用基于 MS-COCO 数据集预训练的 Faster R-CNN 模型检测每个关键帧中的人物, 得到人物边界框; 最后, 考虑到画面中可能存在多个人物的情况, 使用预训练的 TalkNet[30] 识别说话人, 得到说话人边界框. 本文结合说话人边界框 B 和由 Faster R-CNN 提取的特征表示 $\boldsymbol{f}$,

① https://pypi.org/project/scenedetect/

得到视频初始特征 $\boldsymbol{z}^v \in \mathbb{R}^{l_v \times F_v}$, l_v 为关键帧的序列长度, F_v 为每帧的特征维度, $\boldsymbol{z}^v$ 计算为:

$$\boldsymbol{z}^v = AvgPool(RoIAlign(\boldsymbol{f}, B)), \tag{1}$$

其中 *RoIAlign* 表示根据边界框 B 抽取固定大小的特征图, *AvgPool* 用来固定长宽到统一的大小.

2.2 多模态表示

大部分多模态数据存在互补性和一致性. 例如, 人在表达情感或意图时, 表情、语音和语言具有共同的动机和目标, 说明模态间具有一致性的共享特征. 同时, 表情、语音和语言又分别具备特有的情感、语气和语义, 说明模态间具有互补性的特有特征. 所以, 本文设计模态特定和模态共用的 2 类编码器, 分别学习文本、语音、视频的特有特征和共享特征, 为多模态学习提供一个全面的表征视图. 文本、语音和视频的初始特征 $\boldsymbol{z}^t$, $\boldsymbol{z}^a$, $\boldsymbol{z}^v$ 输入编码器之前, 先进行 L2 归一化, 再通过不同的 Transformer 进行预处理, 然后对 Transformer 的输出序列进行累加求平均, 分别得到 $\boldsymbol{u}^t \in \mathbb{R}^{d_t}$, $\boldsymbol{u}^a \in \mathbb{R}^{d_t}$, $\boldsymbol{u}^v \in \mathbb{R}^{d_t}$, d_t 为 Transformer 的最后一层前馈神经网络的输出维度.

2.2.1 共享特征和特有特征表示

1)共享特征表示. 为了学习不同模态的共享特征表示, 构建一个模态共用的编码器 $E_c(\boldsymbol{u}^{(t,v,a)};\theta_c)$, 把文本特征 $\boldsymbol{u}^t$、视觉特征 $\boldsymbol{u}^v$ 和语音特征 $\boldsymbol{u}^a$ 映射到同一个特征空间, 分别得到文本、视频和语音的共享特征 $\boldsymbol{h}_c^t \in \mathbb{R}^{d_c}$, $\boldsymbol{h}_c^v \in \mathbb{R}^{d_c}$, $\boldsymbol{h}_c^a \in \mathbb{R}^{d_c}$, 如式(2)~(4)所示, 其中 θ_c 和 d_c 分别为共用编码器的参数和输出维度.

$$\boldsymbol{h}_c^t = E_c(\boldsymbol{u}^t;\theta_c), \tag{2}$$

$$\boldsymbol{h}_c^v = E_c(\boldsymbol{u}^v;\theta_c), \tag{3}$$

$$\boldsymbol{h}_c^a = E_c(\boldsymbol{u}^a;\theta_c). \tag{4}$$

2)特有特征表示. 为了学习不同模态的特有特征表示, 分别为文本、语音、视频构建一个特定的编码器 $E_p(\boldsymbol{u}^t;\theta_p^t)$, $E_p(\boldsymbol{u}^a;\theta_p^a)$, $E_p(\boldsymbol{u}^v;\theta_p^v)$, 把文本特征 $\boldsymbol{u}^t$、语音特征 $\boldsymbol{u}^a$ 和视觉特征 $\boldsymbol{u}^v$ 映射到不同的特征空间, 分别得到文本、语音和视频的特有特征 $\boldsymbol{h}_p^t \in \mathbb{R}^{d_p}$, $\boldsymbol{h}_p^a \in \mathbb{R}^{d_p}$, $\boldsymbol{h}_p^v \in \mathbb{R}^{d_p}$, 如式(5)~(7)所示, 其中 θ_p^t, θ_p^a, θ_p^v 为特定编码器的参数, d_p 为特定编码器的输出维度, 输出维度和共用编码器的一致.

$$\boldsymbol{h}_p^t = E_p(\boldsymbol{u}^t;\theta_p^t), \tag{5}$$

$$\boldsymbol{h}_p^a = E_p(\boldsymbol{u}^a;\theta_p^a), \tag{6}$$

$$\boldsymbol{h}_p^v = E_p(\boldsymbol{u}^v;\theta_p^v). \tag{7}$$

2.2.2 特征表示学习损失函数

1)样本内协同损失函数

在同一个样本内, 需要保证不同模态的共享特征具有相似性和特有特征具有差异性, 同一模态的共享特征和特有特征具有差异性. 本文利用中心矩差异(central moment discrepancy, *CMD*)和正交性衡量特征之间的相似性和差异性. *CMD* 通过匹配 2 个表示的顺序矩差来计算它们之间的差异, 相比于 KL 散度包含了高阶矩信息, 相比于最大平均差异(maximum mean discrepancy, *MMD*)则减少了计算量, 因为不需要计算核矩阵. 令 $\tilde{X}$ 和 $\tilde{Y}$ 为区间 $[a, b]^N$ 上分别具有概率分布 p 和 q 的有界随机样本, 中心矩差异正则化项 CMD_K 被定义为 *CMD* 的经验估计, 其计算如式(8)所示:

$$CMD_K(\tilde{X},\tilde{Y}) = \frac{1}{|b-a|}\|E(\tilde{X}) - E(\tilde{Y})\|_2 + \sum_{k=2}^{K}\frac{1}{|b-a|^k}\|C_k(\tilde{X}) - C_k(\tilde{Y})\|_2, \tag{8}$$

其中 $E(\tilde{X}) = \frac{1}{|\tilde{X}|}\sum_{x\in X} x$ 表示样本 $\tilde{X}$ 的经验期望向量, $C_k(\tilde{X}) = E\left(\prod_{i=1}^{N}(x_i - E(\tilde{X}))^k\right)$ 表示 $\tilde{X}$ 的 k 阶样本中心距向量. 直观上理解, 如果样本 $\tilde{X}$ 和 $\tilde{Y}$ 的概率分布越相似, 那么它们的每阶中心距也越相近, *CMD* 值越小. 利用 *CMD* 构建共享特征相似度损失函数, 如式(9)所示; 利用正交性约束构建特有特征差异性损失函数, 如式(10)所示; 样本内多模态协同的总损失函数如式(11)所示.

$$L_{intra}^{sim} = \frac{1}{3}\sum_{\substack{(m_1,m_2)\in\\ \{(t,a),(t,v),\\ (a,v)\}}} CMD_K(\boldsymbol{h}_c^{m_1}, \boldsymbol{h}_c^{m_2}), \tag{9}$$

$$L_{intra}^{diff} = \frac{1}{6}\left(\sum_{m\in\{t,a,v\}}\cos(\boldsymbol{h}_p^m, \boldsymbol{h}_c^m) + \sum_{\substack{(m_1,m_2)\in\\ \{(t,a),(t,v),\\ (a,v)\}}}\cos(\boldsymbol{h}_p^{m_1}, \boldsymbol{h}_p^{m_2})\right), \tag{10}$$

$$L_{intra} = L_{intra}^{sim} + L_{intra}^{diff}, \tag{11}$$

其中 t, a, v 分别表示文本、语音和视频, $\boldsymbol{h}_c^m$ 和 $\boldsymbol{h}_p^m$ 分别表示模态 m 的共享特征和特有特征. $\boldsymbol{h}_c^{m_1}$ 和 $\boldsymbol{h}_c^{m_2}$ 越相似 L_{intra}^{sim} 值越小, $\boldsymbol{h}_p^m$ 和 $\boldsymbol{h}_c^m$、$\boldsymbol{h}_p^{m_1}$ 和 $\boldsymbol{h}_p^{m_2}$ 相差越大 L_{intra}^{diff} 值越小.

2)样本间协同损失函数

在不同样本间, 需要保证同类别样本的特征具有相似性和不同类别样本的特征具有差异性. 借鉴对比学习的思路, 在一组样本中随机选择一个样本作为锚点样本 *s*, 与 *s* 类别相同的 N 个样本作为正样本 *pos*, 与 *s* 类别不同的 M 个样本作为负样本 *neg*. 基于 *CMD*, 构建如式(12)所示的样本间多模态协同的

损失函数：

$$L_{\text{inter}}=\frac{1}{6}\sum_{n\in\{\text{c,p}\}}\sum_{m\in\{t,a,v\}}\left(\frac{1}{N}\sum_{i=1}^{N}CMD_K(s(\boldsymbol{h}_n^m),pos_i(\boldsymbol{h}_n^m))-\frac{1}{M}\sum_{j=1}^{M}CMD_K(s(\boldsymbol{h}_n^m),neg_j(\boldsymbol{h}_n^m))\right),\tag{12}$$

其中 c 和 p 分别为共享和特有特征的标识，t, a, v 分别表示文本、语音和视频，$s(\boldsymbol{h}_n^m)$, $pos_i(\boldsymbol{h}_n^m)$, $neg_j(\boldsymbol{h}_n^m)$分别表示锚点样本 s、正样本 i、负样本 j 的 m（文本、语音和视频）模态的 n（共享、特有）特征的表示. 正样本 i 和锚点样本 s 越相似、负样本 j 和锚点样本 s 差异性越大，损失函数 L_{inter} 值越小.

3）样本重构损失函数

为了保证由编码器得到的共享特征和特有特征保留了初始特征空间的相关性质，设计一个解码器 $D(\boldsymbol{h}_{\text{c}}^m,\boldsymbol{h}_{\text{p}}^m;\theta_{\text{d}})$，输入模态 m 的共享特征 $\boldsymbol{h}_{\text{c}}^m$ 和特有特征 $\boldsymbol{h}_{\text{p}}^m$，希望输出能够重构该模态的初始特征. 本文使用均方误差（mean squared error, *MSE*）衡量重构误差，计算如式（13）所示：

$$L_{\text{recon}}=\frac{1}{3}\sum_{m\in\{t,a,v\}}\|\boldsymbol{u}^m-D(\boldsymbol{h}_{\text{c}}^m,\boldsymbol{h}_{\text{p}}^m)\|_2+\frac{\lambda}{2}\|\theta_{\text{d}}\|_2,\tag{13}$$

其中 $\boldsymbol{u}^m$ 表示模态 m 的初始特征表示，$D(\boldsymbol{h}_{\text{c}}^m,\boldsymbol{h}_{\text{p}}^m)$ 为解码器的输出，θ_{d} 为解码器的参数，$\frac{\lambda}{2}\|\theta_{\text{d}}\|_2$ 为正则化项，用于防止过拟合.

2.3 多模态融合

在多模态自然语言分析任务中，文本为主要特征，语音和视频为辅助特征，并且在某些时刻语音和视频包含噪声数据，对结果的判断起到干扰作用. 因此，本文设计一种基于注意力机制和门控神经网络的自适应融合方法.

2.3.1 模态内共享特征和特有特征融合

对每个模态的共享特征和特有特征进行拼接，输入 Self-attention 中，捕获共享特征和特有特征的相关性，得到单模态融合特征. Self-attention 是 Transformer 的核心组件，相比 RNN 网络结构，其最大的优点是可以实现并行计算和长距离依赖. 其计算流程如图 3 所示.

计算形式如式（14）所示：

$$SA(\boldsymbol{X})=softmax\left(\frac{(\boldsymbol{W}^{\text{q}}\boldsymbol{X})(\boldsymbol{W}^{\text{k}}\boldsymbol{X})^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)(\boldsymbol{W}^{\text{v}}\boldsymbol{X}),\tag{14}$$

其中 $\boldsymbol{Q}=\boldsymbol{W}^{\text{q}}\boldsymbol{X}$, $\boldsymbol{K}=\boldsymbol{W}^{\text{k}}\boldsymbol{X}$, $\boldsymbol{V}=\boldsymbol{W}^{\text{v}}\boldsymbol{X}$ 分别为 Query, Key, Value 矩阵，$\boldsymbol{W}^{\text{q}}$, $\boldsymbol{W}^{\text{k}}$, $\boldsymbol{W}^{\text{v}}$ 为需要学习的权重矩阵. d_{k} 为 Key 的维度，除以 $\sqrt{d_{\text{k}}}$ 的目的是在反向传播时梯度更加稳定. Self-attention 的 Query, Key, Value 来自于同一个输入序列 $\boldsymbol{X}$. 分别对文本、语音和视频的共享特征和特有特征进行拼接 $\boldsymbol{X}=Concat(\boldsymbol{h}_{\text{c}}^m,\boldsymbol{h}_{\text{p}}^m)$，输入 Self-attention 中得到单模态融合特征 $\boldsymbol{h}^t\in\mathbb{R}^{d_{\text{v}}}$, $\boldsymbol{h}^a\in\mathbb{R}^{d_{\text{v}}}$, $\boldsymbol{h}^v\in\mathbb{R}^{d_{\text{v}}}$, d_{v} 为 Value 的维度.

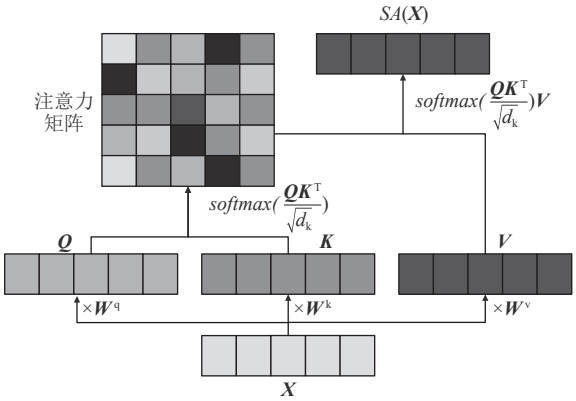

Fig. 3 Self-attention calculation process

图 3 Self-attention 计算流程图

2.3.2 样本内多模态特征融合

得到单模态的融合特征后，基于 Cross-attention 分别计算文本与视频的关联特征 $CA(t,v)$ 和文本与语音的关联特征 $CA(t,a)$. 不同于 Self-attention, Cross-attention 的 Query, Key, Value 的输入来自于 2 个不同的序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$, $\boldsymbol{X}$ 作为 Query 的输入，而 $\boldsymbol{Y}$ 作为 Key 和 Value 的输入.Cross-attention 的计算流程如图 4 所示.

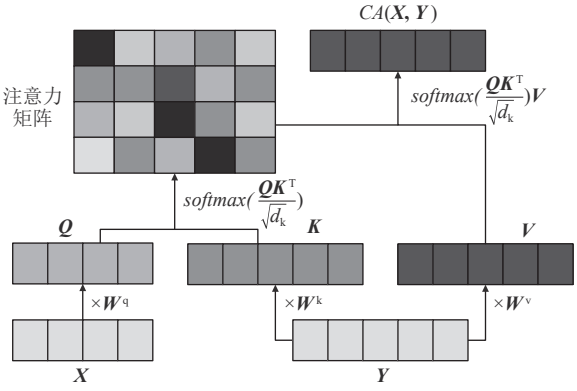

Fig. 4 Cross-attention calculation process

图 4 Cross-attention 计算流程图

计算形式如式（15）所示：

$$CA(\boldsymbol{X},\boldsymbol{Y})=softmax\left(\frac{(\boldsymbol{W}^{\text{q}}\boldsymbol{X})(\boldsymbol{W}^{\text{k}}\boldsymbol{Y})^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)(\boldsymbol{W}^{\text{v}}\boldsymbol{Y}).\tag{15}$$

然后，把 $CA(t,v)$ 和 $CA(t,a)$ 分别输入视觉门控神经单元和语音门控神经单元，得到视觉特征融合权重 w^v 和语音特征融合权重 w^a. 最后，根据权重融合视觉特征 $\boldsymbol{h}^v$、语音特征 $\boldsymbol{h}^a$ 和文本特征 $\boldsymbol{h}^t$，得到最终的多模态融合特征 $\boldsymbol{h}$，如式（16）所示：

$$\boldsymbol{h}=\boldsymbol{h}^t+w^v\boldsymbol{h}^v+w^a\boldsymbol{h}^a. \tag{16}$$

直观上理解, w^v和w^a根据模态间的深层关系得到, 当视觉特征和语音特征能辅助文本特征做决策时, 则增加融合权重, 反之则减少. 这种融合方式, 一方面体现了文本特征为主、语音和视觉特征为辅的先验; 另一方面, 实现了自适应的融合, 有效降低了语音和视频中可能存在的噪声干扰.

2.4 结果预测

把多模态融合特征 $\boldsymbol{h}$ 输入多层全连接神经网络中进行分类或回归任务. 分类任务使用交叉熵损失, 回归任务使用均方误差损失, 如式(17)所示:

$$L_{\text{task}}=\begin{cases}-\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}y_i\ln(\hat{y}_i)+\dfrac{\lambda}{2}\|\boldsymbol{W}\|_2, & \text{分类},\\ -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\|y_i-\hat{y}_i\|_2+\dfrac{\lambda}{2}\|\boldsymbol{W}\|_2, & \text{回归},\end{cases} \tag{17}$$

其中 N 是训练样本数量, y_i和$\hat{y}_i$分别代表样本 i 的真实值和预测值, $\dfrac{\lambda}{2}\|\boldsymbol{W}\|_2$为 L2 正则化, 以降低模型的过拟合程度. 为了实现多模态表示、融合和预测端到端的训练, 本文对多模态表示学习损失 L_{intra}、L_{inter}、L_{recon}和预测结果损失L_{task}进行联合优化, 最终优化目标如式(18)所示, 其中α, β, γ是权重.

$$L=L_{\text{task}}+\frac{1}{N}\sum_{i=1}^{N}(\alpha L_{\text{intra}}^i+\beta L_{\text{inter}}^i+\gamma L_{\text{recon}}^i). \tag{18}$$

3 实验与分析

3.1 实验设置

3.1.1 实验数据

本文选取多模态意图识别数据集 MIntRec[①]和多模态情感数据集 CMU-MOSI, CMU-MOSEI[②]作为实验数据, 这 3 个数据集都包含文本、语音和视频 3 种模态.MIntRec 数据集由清华大学智能技术与系统国家重点实验室提供, 原始数据来源于美剧“Superstore”, 包含 2 224 条实例.MIntRec 数据集包含“表达情绪或态度”和“实现目标”2 个粗粒度类别. “表达情绪或态度”细分为 11 个意图类别: Complain, Praise, Apologize, Thank, Criticize, Care, Agree, Taunt, Flaunt, Oppose, Joke. “实现目标”细分为 9 个意图类别: Inform, Advise, Arrange, Introduce, Comfort, Leave, Prevent, Greet, Ask for help. CMU-MOSI 和 CMU-MOSEI 数据集由卡梅隆大学提供, 原始数据来源于 YouTube, 包含强烈积极(+3)、积极(+2)、弱积极(+1)、中性(0)、弱消极(–1)、消极(–2)、强烈消极(–3)这 7 种情感类别. CMU-MOSI 数据集收录了 89 位 YouTube 用户的 2 199 条视频片段, CMU-MOSEI 数据集是 CMU-MOSI 的扩展版, 收录了 1 000 位 YouTube 用户的 3 228 条视频, 包括 250 个主题, 共 23 453 个句子. 训练集、验证集和测试集的划分结果如表 1 所示.

Table 1 Division Results of Datasets

表 1 数据集划分结果

数据集	训练集	验证集	测试集
MIntRec	1 344	455	455
CMU-MOSI	1 319	440	440
CMU-MOSEI	16 265	1869	4 643

3.1.2 评价指标

在 MIntRec 数据集上执行 20-class 分类任务, 利用准确率(accuracy, Acc)、宏平均精确度(macro precision, MP)、宏平均召回率(macro recall, MR)和宏平均 $F1$-score(macro $F1$-score, MF1)作为算法性能评价指标. 在 CMU-MOSI 和 CMU-MOSEI 数据集上执行回归和分类任务, 回归任务利用平均绝对误差(mean absolute error, MAE)和皮尔逊相关系数(Pearson correlation coefficient, PCC)作为评价指标, 分类任务利用二分类准确率(Acc-2)、$F1$-score 和七分类准确率(Acc-7)作为评价指标.Acc, MP, MR, MF1, PCC, $F1$-score 值越大越好, MAE 值越低越好. 在以往的研究中, CMU-MOSI 和 CMU-MOSEI 数据集根据情感分数有(负, 非负)和(负, 正)2 种二分类做法.

3.1.3 实现细节

在初始特征提取中, 分别基于预训练的 BERT-base-uncased, Wav2vec 2.0, Faster R-CNN 分别提取文本、语音和视觉特征. 在多模态表示中, 使用层数为 1、多头个数为 1 的 Transformer 分别对文本、语音和视觉特征进行预处理, 编码器 $E_{\text{c}}(\boldsymbol{u}^{(t,v,a)};\theta_{\text{c}})$, $E_{\text{p}}(\boldsymbol{u}^t;\theta_{\text{p}}^t)$, $E_{\text{p}}(\boldsymbol{u}^v;\theta_{\text{p}}^v)$, $E_{\text{p}}(\boldsymbol{u}^a;\theta_{\text{p}}^a)$和解码器 $D(\boldsymbol{h}_{\text{c}}^m,\boldsymbol{h}_{\text{p}}^m;\theta_{\text{d}})$都采用单层全连接神经网络. 在多模态融合中, Self-attention 和 Cross-attention 采用的层数和多头个数都为 1.在结果预测中, 采用 2 层的全连接神经网络. 整体来说, 考虑到实验数据集相对较小, 没有采用较为深层的网络架构. 其他超参数如表 2 所示. 为了减少式(18)中

① https://drive.google.com/drive/folders/18iLqmUYDDOwIiiRbgwLpzw76BD62PK0p?usp=sharing

② https://github.com/pliang279/MultiBench

超参数α, β, γ的搜索时间，本文采用了一种次优网格搜索方法，具体内容见 3.2.1 节.

Table 2 Hyperparameters Setting

表 2 超参数设置

参数类型	参数	参数值
模型参数	文本的最大序列长度l_t	30
	语音的最大序列长度l_a	230
	视频的最大序列长度l_v	480
	文本特征维度F_t	768
	语音特征维度F_a	256
	视觉特征维度F_v	768
	L_{intra}的权重 α	0.7
	L_{inter}的权重 β	0.7
	L_{recon}的权重 γ	0.6
	中心矩差异 CMD_K 的阶数 K	5
	全连接神经网络的隐藏层大小	256
训练参数	学习率	3E–5
	正则化参数	1E–6
	最大训练轮次	20
	停止训练的等待次数	6
	批量训练的 batch size	8
	Dropout 比例	0.1
	优化器 Optimizer	Adam

3.2 结果和分析

3.2.1 联合优化的超参数分析

L_{intra}, L_{inter}, L_{recon}的权重值α, β, γ是重要的超参数. 我们设定这些参数的搜索空间为{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. 然而，我们没有对α, β, γ同时建立网格，而是先固定其中 2 个参数，只对其中 1 个参数进行搜索，虽然这种网格搜索方式可能得不到最佳的参数组合，但可以极大地减少搜索的时间消耗. 为了提升模型的泛化能力，我们并没有针对不同的数据集选择不同的参数组合. 相反，我们选择了在 MIntRec, CMU-MOSI, CMU-MOSEI 这 3 个数据上平均准确率最高的参数组合. 这样的做法旨在确保模型在不同数据集上都能取得较好的性能，而不仅仅局限于某个特定数据集. 实验结果如图 5 所示.

首先，我们将β和γ固定为 0.5, α=0.7 时模型表现最佳；然后，我们将α固定为 0.7, γ固定为 0.5, β=0.7 时模型表现最佳；最后，我们将α和β固定为 0.7, γ=0.6 时模型的性能最佳. 所以，最终选择α=0.7, β=0.7, γ=0.6.从图 5 看出，模型对重构误差L_{recon}的权重γ更加敏感.

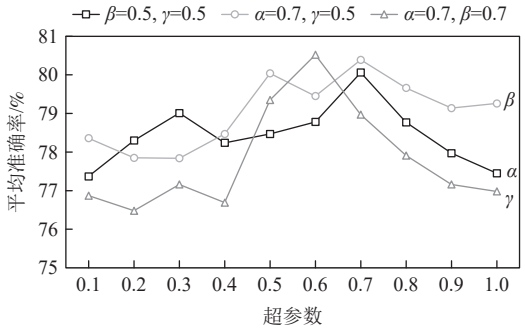

Fig. 5 Hyperparameter search

图 5 超参数搜索

3.2.2 方法对比分析

1）多模态情感分析任务的实验对比

为了验证本文提出的 CoAdMu 方法的有效性，选择以下多模态学习方法作为基线，对比其在多模态情感分析任务中的性能.

①TFN[15]. 一种基于张量的多模态融合方法，对提取的语言特征、视觉特征和语音特征做外积，得到融合的向量.

②LMF[16]. 通过将张量和权重并行分解，利用模态特定的低阶因子来执行多模态融合，避免计算高维的张量.

③MFM[7]. 通过模态分解将多模态表征分解为多模态判别因子和特定模态生成因子，多模态判别因子在所有模态之间共享，特定模态生成因子对于每个模态都是唯一的.

④RMFN[22]. 一种基于循环多阶段融合网络的多模态融合方法，将融合分解成前后关联的多个阶段.

⑤CIA[31]. 采用自编码器学习模态之间的交互关系，并利用上下文感知注意力学习相邻话语间的关系.

⑥MCTN[8]. 通过 Seq2Seq 模型实现不同模态之间的来回转换，得到多模态间的联合表示.

⑦RAVEN[9]. 通过视觉特征和语音特征动态调整文本中词嵌入，实现文本和非文本特征的联合表示.

⑧MulT[23]. 利用跨模态 Transformer 将源模态转换为目标模态来学习多模态表示.

⑨ICCN[6]. 使用语音-文本和视频-文本的特征外积和深度典型相关分析来生成多模态特征表示.

⑩MISA[3]. 将每个模态投射到 2 个不同的子空间中，分别学习共享特征和特有特征，但该方法只考虑了单个样本内的多模态协同.

⑪MAG-BERT[25]. 通过多模态适应门，允许 BERT 在微调期间接受多模态非语言数据，使得语言模型 BERT 有效利用了语音和视觉模态的信息.

⑫QMF[32]. 利用量子理论中的叠加和纠缠来表述单模态和跨模态的相互作用，提高多模态融合的可解释性.

⑬HyCon[10]. 通过联合模态内对比学习、模态间对比学习和半对比学习实现多模态表示.

⑭EMFRM[11]. 在 MISA 方法的基础上，通过中心矩差异对模态的特征空间进行约束，但该方法也只考虑了单个样本内的多模态协同.

表 3 和表 4 分别是在多模态情感数据集 CMU-MOSI 和 CMU-MOSEI 上的对比结果. 大部分基于表示学习的多模态方法的性能优于基于张量融合的多模态方法，因为基于张量融合的方法计算维度呈指数级增长，导致计算效率差，需要大量的训练数据才能学到模态间的交互. 表示学习方法一般基于对多模态数据的先验知识，例如多模态数据的一致性和互补性，构建表示学习模型，降低了学习的复杂度. 在表示学习方法中，基于协同表示的多模态方法的性能优于大部分基于联合表示的方法，因为协同表示方法把多模态信息映射到不同的特征空间，相比于映射到同一个空间的联合表示更好地保留了模态的特有特征. 联合表示方法 MAG-BERT 能获得相对较好的性能，主要是因为其借助了预训练语言模型 BERT 强大的语义学习能力，巧妙地把语音和视觉信息集成到了 BERT 之中.QMF 方法利用量子理论中的叠加和纠缠来表述单模态和跨模态的交互，虽然方法性能没有得到很大的提升，但提高了多模态融合的可解释性. 本文提出的 CoAdMu 方法在 CMU-MOSI 和 CMU-MOSEI 这 2 个数据集上，比最先进的基线方法在所有的评价指标都有一定程度的提升，证明了本文方法的有效性.

2）多模态意图识别任务的实验对比

为了验证 CoAdMu 方法在复杂多模态场景下的学习能力，本文在真实世界多模态意图识别的基准数据集 MIntRec 上，实施进一步的实验对比. 一方面，MIntRec 原始数据来源于真实世界的影视片段，具有丰富的人物角色和故事情节，以及复杂的场景画面；另一方面，MIntRec 包含了更为细粒度的 20 个意图类别，囊括了表达情绪和态度、实现目标. 本文挑选 MulT, MISA, MAG-BERT, HyCon 这 4 个先进的多模态学习方法作为基线. 整体实验结果对比如表 5 所示，每个类别的 *F*1-score 如表 6 和表 7 所示.

从表 5 中可以发现，CoAdMu 方法在所有的评价指标上获得了最好的效果，比最先进的基线方法在 Acc, MP, MR, MF1 上分别提高了 1.35 个百分点、2.47 个百分点、2.35 个百分点、2.22 个百分点，提升效果比在多模态情感分析任务上更加明显. 一方面，MIntRec 数据集在真实场景下采集，可能存在不同级别的噪声；另一方面，MIntRec 数据集中的每种模态

Table 3 Comparison with Baselines on CMU-MOSI Dataset

表 3 在 CMU-MOSI 数据集上与基线的对比

方法	MAE	PCC	Acc-2/%	*F*1-score/%	Acc-7/%
TFN	0.970	0.633	73.9/–	73.4/–	32.1
LMF	0.912	0.668	76.4/–	75.7/–	32.8
MFM	0.951	0.662	78.1/–	78.1/–	36.2
RMFN	0.922	0.681	78.4/–	78.0/–	38.3
CIA	0.914	0.689	79.9/–	79.5/–	38.9
MCTN	0.909	0.676	79.3/–	79.1/–	35.6
RAVEN	0.915	0.691	78.0/–	76.6/–	33.2
MulT	0.871	0.698	–/83.0	–/82.8	40.0
ICCN	0.862	0.714	–/83.0	–/83.0	39.0
MISA	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG-BERT	0.790	0.769	82.2/83.5	82.6/83.5	42.9
QMF	0.915	0.696	–/79.7	–/79.6	33.5
HyCon	0.713	0.790	–/85.2	–/85.1	46.6
EMRFM	0.722	0.785	–/84.7	–/84.8	46.1
CoAdMu（本文）	**0.711**	**0.798**	**84.1/86.1**	**84.0/86.1**	**47.2**
Δ_{SOTA}	↓0.002	↑0.008	↑1.9/↑0.9	↑1.4/↑1.0	↑0.6

注：Δ_{SOTA} 表示本文方法和最先进的方法对比，↓表示下降，↑表示提升，–/–的左右侧分别代表（负，非负）和（负，正）的结果. 黑体数值表示最优值.

Table 4 Comparison with Baselines on CMU-MOSEI Dataset

表 4 在 CMU-MOSEI 数据集上与基线的对比

方法	MAE	PCC	Acc-2/%	*F*1-score/%	Acc-7/%
TFN	0.610	0.671	79.4/–	79.7/–	49.8
LMF	0.608	0.677	80.6/–	81.0/–	50.0
MFM	0.602	0.692	81.1/–	81.6/–	50.7
RMFN	0.604	0.685	80.9/–	81.2/–	50.5
CIA	0.680	0.590	80.4/–	78.2/–	50.1
MCTN	0.609	0.670	79.8/–	80.6/–	49.6
RAVEN	0.614	0.662	79.1/–	79.5/–	50.0
MulT	0.580	0.703	–/82.5	–/82.3	51.8
ICCN	0.565	0.713	–/84.2	–/84.2	51.6
MISA	0.555	0.756	83.6/85.5	83.8/85.3	52.2
MAG-BERT	0.602	0.778	83.1/85.0	83.2/85.0	51.9
QMF	0.640	0.658	–/80.7	–/79.8	47.9
HyCon	0.601	0.776	–/85.4	–/85.6	52.8
EMRFM	0.600	0.775	–/85.2	–/85.3	52.3
CoAdMu（本文）	**0.550**	**0.791**	**84.2/86.5**	**84.6/86.7**	**53.6**
$\triangle_{\text{SOTA}}$	↓0.005	↑0.013	↑0.6/↑1.0	↑0.8/↑1.1	↑0.8

注：$\triangle_{\text{SOTA}}$ 表示本文方法和最先进的方法对比，↓表示下降，↑表示提升，–/–的左右侧分别代表（负，非负）和（负，正）的结果. 黑体数值表示最优值.

Table 5 Comparison with Baselines on MIntRec Dataset

表 5 在 MIntRec 数据集上与基线的对比 %

方法	Acc	MP	MR	MF1
MulT	71.24	67.53	68.15	67.58
MISA	71.91	69.98	68.91	68.92
MAG-BERT	71.01	68.15	65.83	66.09
HyCon	71.33	68.93	65.21	66.37
CoAdMu（本文）	**73.26**	**72.45**	**71.26**	**71.14**
$\triangle_{\text{SOTA}}$	↑1.35	↑2.47	↑2.35	↑2.22

注：$\triangle_{\text{SOTA}}$ 表示本文方法和最先进的方法对比，↓表示下降，↑表示提升. 黑体数值表示最优值.

在不同时刻可能表现出不同的作用类型. 例如, 在表达 Agree 和 Thank 这 2 种意图时的文本具有相对固定的表达方式, 语音和视觉特征的作用并不明显. 然而, 在表达 Taunt 和 Joke 意图时, 语音和视频中的语气和表情是很好的补充特征. 基线方法基本上都是“重表示, 轻融合”, 在表示学习上设计了复杂的方式, 而在特征融合上采用简单的拼接方式, 没有突出不同特征的作用大小, 导致模型对噪声数据敏感.CoAdMu 采用协同表示和自适应融合的方式, 不仅很好地学习了多模态表示, 而且可以根据不同的类别类型、作用大小自动调整特征的融合权重, 有效降低了噪声数据的干扰, 所以获得了相对较好的效果.

从表 6 和表 7 发现, 同一个方法在不同的意图分类上具有不同的性能, 没有哪个方法能够在所有的类别上获得最好的性能.MulT 在 Complain, Joke, Greet 这 3 个类别上获得了最高评分; MISA 在 Agree, Flaunt, Arrange, Introduce, Prevent 这 5 个类别上获得了最高评分; MAG-BERT 在 Apologize, Oppose, Inform, Ask for help 这 4 个类别上获得了最高评分;

Table 6 *F*1-score for Each Fine-grained Intent Category in “Express Emotions and Attitudes”

表 6 “表达情绪或态度”中每个细粒度意图类别的 *F*1-score %

方法	Complain	Praise	Apologize	Thank	Criticize	Care	Agree	Taunt	Flaunt	Joke	Oppose
MulT	**67.26**	87.36	98.11	95.83	48.00	86.49	91.67	9.52	36.36	**50.00**	35.29
MISA	62.14	86.67	98.11	98.04	47.06	81.82	**100.00**	25.00	**50.00**	37.50	27.27
MAG-BERT	66.09	90.24	**98.18**	98.04	40.00	85.71	95.65	9.09	15.38	40.00	**40.00**
HyCon	66.67	**93.83**	98.11	**98.11**	48.89	**94.74**	95.65	10.00	23.53	28.57	38.10
CoAdMu（本文）	64.08	90.24	96.30	94.34	**60.87**	91.89	96.00	**28.57**	44.44	40.00	34.78

注：黑体数值表示最优值.

Table 7 *F*1-score for Each Fine-grained Intent Category in "Achieve Goals"

表 7 "实现目标"中每个细粒度意图类别的 *F*1-score %

方法	Comfort	Inform	Advise	Arrange	Introduce	Leave	Prevent	Greet	Ask for help
MulT	68.57	69.72	72.00	64.00	60.00	70.27	80.00	**91.67**	69.57
MISA	78.79	69.57	70.59	**65.00**	**72.00**	75.00	**85.71**	90.91	57.14
MAG-BERT	70.00	**70.23**	60.47	62.22	61.90	75.00	80.00	90.91	**72.73**
HyCon	74.29	65.67	**72.73**	62.50	59.46	68.97	69.23	85.71	72.73
CoAdMu（本文）	**83.33**	69.64	77.19	57.78	71.79	**83.87**	82.76	85.71	70.59

注：黑体数值表示最优值.

HyCon 在 Praise, Thank, Care, Advise 这 4 个类别上获得了最高评分; CoAdMu 在 Criticize, Taunt, Comfort, Leave 这 4 个类别上获得了最高评分. 虽然 CoAdMu 获得的最高评分总个数不是最多, 但平均的 *F*1-score 值最大. 通过最高分的分布发现, HyCon 擅长于情感表达类的意图识别, MISA 擅长于表达态度和实现目标类的意图识别, MulT, MAG-BERT, CoAdMu 在不同的任务上表现比较均衡. 所有方法在大部分意图类别上都能获得较好的分类效果, 但在 Taunt, Flaunt, Joke, Oppose 这 4 个类别上的分类效果不佳, 因为这些意图的识别需要结合语言、语气、表情、动作和情景等做深层次的推理, 这也说明 CoAdMu 和基线方法在多模态深层推理任务上还存在不足.

3.2.3 消融实验分析

为了进一步分析不同模块对 CoAdMu 的贡献, 我们设计了 11 组消融实验方法. 方法①～③是对不同模态的消融; 方法④～⑥是对多模态表示模块中相关损失函数的消融; 方法⑦去除多模态表示模块, 直接把 $\boldsymbol{u}^t$, $\boldsymbol{u}^a$, $\boldsymbol{u}^v$ 输入多模态融合层; 方法⑧采用分段训练方式, 首先根据损失函数 $L = L_{\text{intra}} + L_{\text{inter}} + L_{\text{recon}}$ 对多模态表示模块进行预训练, 然后, 再根据损失函数 L_{task} 对多模态融合模块和结果预测模块进行训练; 方法⑨⑩分别去除多模态特征融合中的视觉门控神经单元和语音门控神经单元; 方法⑪去除整个多模态融合模块, 采用简单相加的方式进行融合. 实验结果如表 8 所示.

Table 8 Ablation Experiment Results

表 8 消融实验结果

方法	CMU-MOSI		CMU-MOSEI		MIntRec	
	MAE	Acc-7/%	MAE	Acc-7/%	Acc/%	MF1/%
CoAdMu（本文）	0.711	47.2	0.550	53.6	73.3	71.1
①（–）Text	1.372	22.6	0.790	24.6	29.9	22.8
②（–）Video	0.786	44.7	0.561	50.9	71.5	67.4
③（–）Audio	0.730	46.1	0.557	52.0	73.1	70.8
④（–）L_{intra}	0.791	43.9	0.568	49.4	71.4	68.9
⑤（–）L_{inter}	0.734	45.8	0.559	51.6	72.6	70.5
⑥（–）L_{recon}	0.798	43.2	0.570	49.1	69.5	67.6
⑦（–）MultRe	0.803	42.6	0.575	48.4	70.6	68.5
⑧（*）MultRe	0.784	44.9	0.562	50.9	71.6	69.4
⑨（–）Gate_V	0.805	42.3	0.572	48.7	69.6	66.7
⑩（–）Gate_A	0.728	46.6	0.555	52.5	72.6	70.1
⑪（–）MultFu	0.807	42.0	0.578	48.0	69.1	66.3

通过方法①～③的实验结果发现, 去除文本对模型性能的影响最大, 一方面的原因是文本相比于语音和视频包含了更多的信息量; 另一方面得益于大规模预训练语言模型的应用, 提取的文本特征的质量远高于语音和视觉特征. 去除视频比去除语音对方法的性能影响更大, 这是因为相比于语音, 视觉特征和文本特征的冗余性相对较小, 可以更好地补充文本特征.

通过方法④～⑥的实验结果发现，去除任何一个多模态表示学习的损失函数都会降低模型的性能，这是因为样本内的协同损失函数L_{intra}保证了共享特征的相似性和特有特征的差异性；样本间的协同损失函数L_{inter}保证了同类别样本的特征具有相似性，不同类别样本的特征具有差异性；样本重构损失函数L_{recon}使得共享特征和特有特征保留了初始特征空间的相关性质，避免学习到不相干的特征表示. 通过方法⑦的实验结果发现，如果去除多模态表示学习模块，会对 CoAdMu 的性能造成较大的影响. 这是因为多模态数据存在互补性和一致性，对其共享特征和特有特征分开学习，能提供更加全面的视图. 通过方法⑧的实验结果发现，采用分段训练的方式会降低 CoAdMu 的性能，因为相比端到端的训练方式，分段训练缺乏灵活性和领域适配能力.

通过方法⑨～⑪的实验结果发现，去除多模态融合中的视觉门控神经网络、语音门控神经网络或者整个融合模块都会影响 CoAdMu 的性能，因为语音和视频可能在不同时刻表现出不同的作用类型和不同级别的噪声，门控神经网络可以根据特征对预测结果的作用大小自动分配融合权重，能有效降低噪声的干扰. 相比于语音门控神经网络，视觉门控神经网络起到了更大的作用，这是因为视觉特征相比于语音特征对预测结果起到了更大的作用，方法②③也印证了这一点. 从总体实验结果看，本文设计的每个模块都发挥着各自的作用，去除任何一个模块都会影响方法的性能，证明了 CoAdMu 设计的合理性.

3.2.4 误差分析

为了对预测结果的误差进行详细分析，本文对多模态意图识别的测试结果混淆矩阵进行可视化，如图 6 所示. 横坐标为预测标签，纵坐标为真实标签，颜色深浅代表预测概率的大小，对角线位置的亮度越高，说明 CoAdMu 在该类别的精准率越高. 总体上，CoAdMu 在大部分意图类别上获得了较高的精准率，例如，在 Praise, Apologize, Thank, Agree, Care 等类别上获得了 90% 以上的精准率. 然而，在 Taunt, Joke, Oppose 这 3 个类别上表现不佳，因为这些意图需要结合语言、语气、表情、动作和情景等做深层次的推理，有时候连人也无法准确地判断. 从图 6 也发现，CoAdMu 容易把 Complain 误判成 Criticize 或 Oppose, Criticize 误判成 Taunt, Taunt 误判成 Joke, Joke 误判成 Flaunt, Inform 误判成 Arrange.这也是容易理解的，因为这些意图类别具有很高的相似性，有时人也会误判. 虽然我们通过样本间的协同，保证同类别样本的特征具有相似性，不同类样本的特征具有差异性，但这些类别的语义十分相似，还是不能很好地将其区分开.

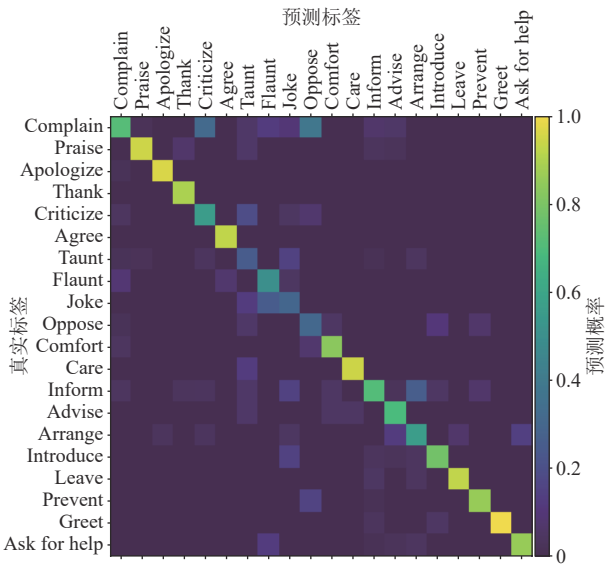

Fig. 6 Visualization of confusion matrix

图 6 混淆矩阵可视化

3.2.5 融合权重分析

为了进一步验证本文提出的多模态特征融合方法的有效性，对 MIntRec 测试集上每个类别的语音和视觉特征的融合权重计算平均值，结果如表 9 所示.

Table 9 Speech and Visual Feature Fusion Weights under Different Intent Categories

表 9 不同意图类别下的语音和视觉特征融合权重

类别	语音特征融合权重	视觉特征融合权重
Advise	0.17	0.97
Agree	0.40	0.98
Apologize	0.93	0.98
Arrange	0.24	0.99
Ask for help	0.15	0.99
Care	0.24	0.97
Comfort	0.80	0.99
Complain	0.92	0.94
Criticize	0.93	0.98
Flaunt	0.88	0.92
Greet	0.47	0.97
Inform	0.73	0.95
Introduce	0.72	0.96
Joke	0.85	0.98
Leave	0.78	0.96
Oppose	0.50	0.99
Praise	0.96	0.68
Prevent	0.52	0.98
Taunt	0.87	0.96
Thank	0.98	0.92

从表 9 中可以看出, 几乎所有类别的视觉特征都获得了很高的权重值, 因为视觉特征可以很好地辅助文本特征, 在消融实验的方法②中也印证了这一点. 而对于语音特征, 不同的类别具有不同的融合权重值, 例如 Apologize, Complain, Criticize, Praise, Thank 获得了较高的权重值, 而 Advise, Arrange, Ask for help, Care 的权重值比较低. 这是因为相比于后者, 前者具有明显的语气特征, 能为意图的判断提供帮助. 从这些结果中可以看出, CoAdMu 实现了自适应的多模态特征融合, 当特征能为决策提供有效信息时则增加融合权重, 反之则减少.

3.2.6 可视化特征分布

本文从 MIntRec 测试集中挑选 Complain 和 Inform 这 2 类样本, 利用 PCA 对样本的初始模态特征、共享和特有特征进行降维并可视化, 如图 7 所示.

图 7(a)是经过 Transformer 预处理后的文本初始特征$\boldsymbol{u}^t$、语音初始特征$\boldsymbol{u}^a$和视频初始特征$\boldsymbol{u}^v$的分布; 图 7(b)是只考虑样本内多模态协同的共享特征$\boldsymbol{h}_{\mathrm{c}}^t$, $\boldsymbol{h}_{\mathrm{c}}^a$, $\boldsymbol{h}_{\mathrm{c}}^v$和特有特征$\boldsymbol{h}_{\mathrm{p}}^t$, $\boldsymbol{h}_{\mathrm{p}}^a$, $\boldsymbol{h}_{\mathrm{p}}^v$的分布; 图 7(c)是同时考虑样本内和样本间多模态协同的共享特征$\boldsymbol{h}_{\mathrm{c}}^t$, $\boldsymbol{h}_{\mathrm{c}}^a$, $\boldsymbol{h}_{\mathrm{c}}^v$和特有特征$\boldsymbol{h}_{\mathrm{p}}^t$, $\boldsymbol{h}_{\mathrm{p}}^a$, $\boldsymbol{h}_{\mathrm{p}}^v$的分布; 图 7(d)(e)分别是对图 7(c)(b)中视频特有特征$\boldsymbol{h}_{\mathrm{p}}^v$的放大视图.

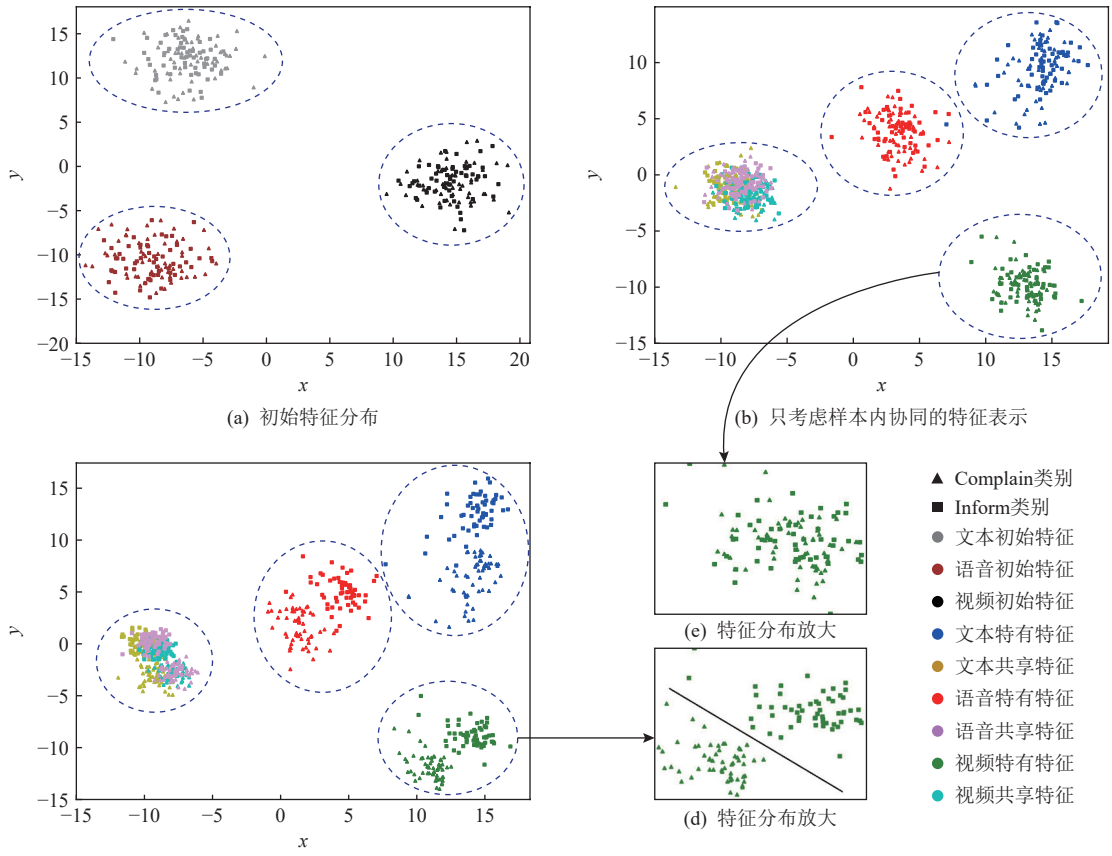

Fig. 7 Feature distribution visualization

图 7 特征分布可视化

从图 7(a)中可以看出, 由于文本、语音和视频的异构性, 它们的初始特征空间的分布具有很大的差异性. 从图 7(b)(e)中可以看出, 只考虑样本内多模态协同, 所有样本的共享特征的分布具有相似性, 特有特征具有差异性, 但不同类别样本的特征没有很好地被区分. 从图 7(c)(d)中可以看出, 同时考虑样本内和样本间的多模态协同, 不仅保证了共享特征的相似性和特有特征的差异性, 而且实现了不同类别样本的特征具有一定的差异性.

4 总 结

多模态表示和融合是多模态机器学习的 2 个关键任务, 针对多模态协同表示时没有考虑样本间协同和多模态融合对噪声数据敏感的问题, 本文提出一种基于样本内外协同表示和自适应融合的多模态

学习方法. 通过构建模态共用和模态特定编码器, 基于样本内和样本间的多模态协同约束, 学习模态的共享特征和特有特征. 通过注意力机制和门控神经网络实现多模态特征的自适应融合, 有效降低了噪声数据的干扰. 在多模态意图识别和多模态情感分析任务上的实验结果表明, 本文方法在多个评价指标上优于基线方法, 大量的消融实验分析、误差分析、融合权重分析和特征可视化分析也证明了本文方法的有效性. 然而, 在实验中也发现本文方法和目前已有方法在多模态联合的深层次推理任务上还有很大的提升空间, 例如讽刺和暗喻识别任务. 在下一步的研究计划中, 将设计融合知识的多模态学习方法, 提升模型在深层次推理任务上的性能.

作者贡献声明: 黄学坚负责研究方案的构建, 并完成实验和撰写论文; 马廷淮提出了算法思路和实验方案; 王根生提出了指导意见并修改论文.

参 考 文 献

[1] Rahate A, Walambe R, Ramanna S, et al. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions[J]. Information Fusion, 2022, 81: 203−239

[2] Liang P P, Lyu Y, Fan Xiang, et al. MultiBench: Multiscale benchmarks for multimodal representation learning[J]. arXiv preprint, arXiv: 2107.07502, 2021

[3] Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 1122−1131

[4] Li Xuelong. Multi-modal cognitive computing[J]. SCIENTIA SINICA Informationis, 2023, 53(1): 1−32(in Chinese)
(李学龙. 多模态认知计算[J]. 中国科学: 信息科学, 2023, 53(1): 1−32)

[5] Wu Yang, Lin Zijie, Zhao Yanyan, et al. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 4730−4738

[6] Sun Zhongkai, Sarma P, Sethares W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 8992−8999

[7] Tsai Y H H, Liang P P, Zadeh A, et al. Learning factorized multimodal representations[J]. arXiv preprint, arXiv: 1806.06176, 2018

[8] Pham H, Liang P P, Manzini T, et al. Found in translation: Learning robust joint representations by cyclic translations between modalities[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6892−6899

[9] Wang Yansen, Shen Ying, Liu Zhun, et al. Words can shift: Dynamically adjusting word representations using nonverbal behaviors[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 7216−7223

[10] Mai Sijie, Zeng Ying, Zheng Shuangjia, et al. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 2276−2289

[11] Huang Xuejian, Ma Tinghuai, Jia Li, et al. An effective multimodal representation and fusion method for multimodal intent recognition[J]. Neurocomputing, 2023, 548: 126373

[12] Zhang Chao, Yang Zichao, He Xiaodong, et al. Multimodal intelligence: Representation learning, information fusion, and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478−93

[13] Zhang Yanyong, Zhang Sha, Zhang Yu, et al. Multi-modality fusion perception and computing in autonomous driving[J]. Journal of Computer Research and Development, 2020, 57(9): 1781−1799(in Chinese)
(张燕咏, 张莎, 张昱, 等. 基于多模态融合的自动驾驶感知及计算[J]. 计算机研究与发展, 2020, 57(9): 1781−1799)

[14] Xu Peng, Zhu Xiatian, Clifton D A. Multimodal learning with transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113−12132

[15] Zadeh A, Chen Minghai, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint, arXiv: 1707.07250, 2017

[16] Liu Zhun, Shen Ying, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint, arXiv: 1806.00064, 2018

[17] Ma Mengmeng, Ren Jia, Zhao Long, et al. SMIL: Multimodal learning with severely missing modality[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 2302−2310

[18] Abdullah S M S A, Ameen S Y A, Sadeeq M A M, et al. Multimodal emotion recognition using deep learning[J]. Journal of Applied Science and Technology Trends, 2021, 2(2): 52−58

[19] Jabeen S, Li Xi, Amin M S, et al. A review on methods and applications in multimodal deep learning[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2s): 1−41

[20] Miyazawa K, Kyuragi Y, Nagai T. Simple and effective multimodal learning based on pre-trained transformer models[J]. IEEE Access, 2022, 10: 29821−29833

[21] Bayoudh K, Knani R, Hamdaoui F, et al. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets[J]. The Visual Computer, 2022, 38: 2939−2970

[22] Liang P P, Liu Ziyin, Zadeh A, et al. Multimodal language analysis with recurrent multistage fusion[J]. arXiv preprint, arXiv: 1808.03920, 2018

[23] Tsai Y H H, Bai Shaojie, Liang P P, et al. Multimodal Transformer for

unaligned multimodal language sequences[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 6558–6569

[24] Mou Luntian, Zhou Chao, Zhao Pengfei, et al. Driver stress detection via multimodal fusion using attention-based CNN-LSTM[J]. Expert Systems with Applications, 2021, 173: 114693

[25] Rahman W, Hasan M K, Lee S, et al. Integrating multimodal information in large pretrained transformers[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 2359−2369

[26] Baevski A, Zhou Yuhao, Mohamed A, et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations[C]// Proc of the 34th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 12449−12460

[27] Hsu W N, Bolte B, Tsai Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451−3460

[28] Chen Sanyuan, Wang Chengyi, Chen Zhengyang, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505−1518

[29] Zhang Hanlei, Xu Hua, Wang Xin, et al. MIntRec: A new dataset for multimodal intent recognition[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 1688−1697

[30] Tao Ruijie, Pan Zexu, Das R K, et al. Is someone speaking?: Exploring long-term temporal features for audio-visual active speaker detection[C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 3927−3935

[31] Chauhan D S, Akhtar M S, Ekbal A, et al. Context-aware interactive attention for multi-modal sentiment and emotion analysis[C]// Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 5647−5657

[32] Li Qiuchi, Gkoumas D, Lioma C, et al. Quantum-inspired multimodal fusion for video sentiment analysis[J]. Information Fusion, 2021, 65: 58−71

Huang Xuejian, born in 1990. PhD candidate, lecturer. Member of CCF. His main research interests include multimodal machine learning and social network analysis.

黄学坚, 1990 年生. 博士研究生，讲师. CCF 会员. 主要研究方向为多模态机器学习、社交网络分析.

Ma Tinghuai, born in 1974. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include social network privacy protection, big data mining, and text emotion computing.

马廷淮, 1974 年生. 博士，教授，博士生导师. CCF 高级会员. 主要研究方向为社交网络隐私保护、大数据挖掘、文本情感计算.

Wang Gensheng, born in 1974. PhD, professor, PhD supervisor. Member of CCF. His main research interests include data mining and social network.

王根生, 1974 年生. 博士，教授，博士生导师. CCF 会员. 主要研究方向为数据挖掘、社交网络.