计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(2): 374−382, 2025

面向情感语义不一致的多模态情感分析方法

罗渊贻　吴　锐　刘家锋　唐降龙
（哈尔滨工业大学计算机科学与技术学院　哈尔滨　150006）
（luoyy@stu.hit.edu.cn）

Multimodal Sentiment Analysis Method for Sentimental Semantic Inconsistency

Luo Yuanyi, Wu Rui, Liu Jiafeng, and Tang Xianglong
（*College of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150006）

Abstract　Multimodal sentiment analysis is a multimodal task that uses multiple modalities of subjective information to analyze sentiment. In some scenarios, the sentimental expression in different modalities is inconsistent, even contradictory, which will weaken the effect of multimodal collaborative decision-making. In this paper, a multimodal learning method is proposed to learn the modal feature representations with consistent sentimental semantics. In order to improve the common feature representation of different modalities and learn the dynamic interaction between modalities without affecting the original information, we first learn the common feature representation of each modality, and then use cross attention to enable one modality to effectively obtain auxiliary information from the common feature representations of other modalities. In multimodal fusion, we propose a multimodal attention, which is used to weighted concatenate modal feature representations, in order to increase the expression of contributed modalities and suppress the influence of weak modalities. The experimental results of the proposed method on the sentiment analysis datasets MOSI, MOSEI, CH-SIMS are better than those of the compared models, indicating the necessity and rationality of considering the problem of sentimental semantic inconsistency in multimodal sentiment analysis.

Key words　multimodal sentiment analysis; common features; sentimental semantic inconsistency; attention mechanism

摘　要　多模态情感分析是利用多种模式的主观信息对情感进行分析判断的一种多模态任务.情感表达具有主观性,在某些场景下不同模态的情感表达不一致,甚至存在相悖的情况,这会削弱多模态协同决策的效果.针对不同模态间情感语义不一致的问题,提出一种多模态学习方法,学习情感语义表达一致的模态特征表示.为了在不影响模态原始信息的同时,提高各模态的共性特征表达并增加模态间的动态交互,首先学习每个模态的共性特征表示,然后利用交叉注意力使单个模态能有效从其余模态的共性特征表示中获取辅助信息.在模态融合模块,以软注意力机制为基础提出模态注意力,对情感语义表达一致的各模态特征表示进行加权连接,以增大强模态的表达,抑制弱模态对任务的影响.提出的模型在情感分析数据集MOSI, MOSEI, CH-SIMS上的实验结果均优于对比模型,表明在多模态情感分析任务中考虑情感语义不一致问题的必要性与合理性.

关键词　多模态情感分析;共性特征;情感语义不一致;注意力机制

中图法分类号　TP18

收稿日期:2023-03-27;**修回日期**:2024-01-23

基金项目:国家自然科学基金项目(61672190)

This work was supported by the National Natural Science Foundation of China (61672190).

通信作者:吴锐(simple@hit.edu.cn)

DOI: 10.7544/issn1000-1239.202330199 CSTR: 32373.14.issn1000-1239.202330199

随着人工智能技术的迅速发展，智能机器逐渐进入大众生活，准确快速地对人类情感进行分析可以使智能机器为人类的日常生活提供更好的服务. 例如在智能汽车系统中，对驾驶员的情感变化进行实时分析，能有效减少意外事故的发生. 人的情感可以通过面部表情、声音等多种模态进行表现，多模态情感分析就是综合多个模态信息对人类情感进行分析决策的一种多模态处理任务[1].

本文的研究对象是以文本、语音、视频3种模态为基础的多模态情感分析任务. 在研究中我们发现，某些场景下不同模态表达的情感语义不一致，甚至相悖. 例如：一个人在悲伤的情况下说“天空真美”，当前场景下的视觉和音频模态表现出消极的情感，而文本表现出积极的情感，此种单模态间语义表达的不一致问题会降低多模态任务分析的准确性. 为直观地从数据方面体现模态情感语义不一致问题对多模态情感分析任务的影响，本文以中文情感数据集CH-SIMS[2]为基础，从2 281个样本中筛选出228个包含模态情感语义不一致的样本，并选取2种现有的多模态情感分析模型：基于深度神经网络的后期融合模型（late fusion based on deep neural network, LF-DNN）和多模态转换器（multimodal transformer, MulT）模型[3]进行对比实验分析，其中LF-DNN模型代码来源于文献[2]提供的可获取开源数据. CH-SIMS数据集包含了多模态标签和单模态标签，有助于分析单模态之间的相关性. 我们将单个样本中某个模态的情感极性与其余任一模态的情感极性相反作为筛选条件，获取情感语义不一致的样本. 分析实验结果如表1所示，其中org表示样本量为2 281的原始数据集，其中训练数据为1 368个样本、验证数据为456个样本、测试数据为457个样本，训练集、验证集、测试集的样本数比例约为3∶1∶1；(–)d表示从原始数据集中筛选出的228条包含情感语义不一致的样本数据；(re)d_{170}表示将170个情感语义不一致的样本以训练集、验证集、测试集的比例3∶1∶1与(–)d中的训练集、验证集、测试集数据进行替换（保证总样本数不变）；(re)d_{228}是将228个情感语义不一致的样本以比例3∶1∶1与(–)d中的训练集、验证集、测试集数据进行替换；(re)d_{test}是将所有情感语义不一致的样本数据与(–)d中的测试集样本进行替换.

表1中的2分类是将情感分为积极和消极2类，范围分别为[1.0, 0]和(0, –1.0]；3分类是将情感分为积极、中性、消极，范围分别为[1.0, 0.2)，[0.2, –0.2]，(–0.2, –1.0]；MAE和Corr分别表示平均绝对误差（mean absolute error, MAE）和皮尔逊相关系数（Pearson correlation coefficient, Corr）. 从表1中可以看出，模态间情感语义不一致的数据会降低多模态情感任务的分析准确度. 在以深度神经网络的后期融合模型LF-DNN为基础的对比实验中，将情感语义不一致的样本数据从原数据集中去除后，实验结果有大幅度提高，但随着将情感语义不一致的样本与其余样本数据逐渐进行替换，实验准确度不断下降，最终在测试集被全部替换后出现了大幅度下跌. 在以多模态转换器模型MulT为基础的对比实验中，表现出的实验现象与LF-DNN相同.

Table 1 Influence of Sample Data with Sentimental Semantic Inconsistency Between Modalities on Multimodal Sentiment Analysis Tasks

表1 模态间情感语义不一致的样本数据对多模态情感分析任务的影响

模型	CH-SIMS数据集					
	数据样本	2分类	3分类	*F*1	MAE	Corr
LF-DNN	org	0.792	0.644	0.802	0.458	0.558
	(–)d	0.830	0.669	0.831	0.408	0.659
	(re)d_{170}	0.804	0.668	0.811	0.418	0.639
	(re)d_{228}	0.793	0.667	0.802	0.432	0.619
	(re)d_{test}	0.699	0.588	0.712	0.530	0.403
MulT	org	0.784	0.647	0.799	0.492	0.555
	(–)d	0.790	0.634	0.807	0.492	0.589
	(re)d_{228}	0.770	0.628	0.790	0.515	0.544

为解决上述问题，本文提出一种多模态情感分析方法以提升模态间情感语义的一致性. 在单模态特征提取方面，现有的多模态情感分析研究只单一地关注模态的时序性或空间性，采用长短期记忆（long short-term memory, LSTM）模型、时序卷积、空间卷积等方法获取模态的时序特征或空间特征表示. 然而视频模态中不仅含有时序信息，而且包含了空间信息[4]，获取视频模态的时空信息能得到更具备表现力的特征信息. 由此，本文利用时间注意力和时空注意力分别获取音频模态和视频模态的特征表示. 据我们所知，目前没有多模态情感分析方面的研究单独采用时空注意力对视频模态进行特征提取. 为解决模态间情感语义不一致的问题，获取语义一致的模态特征表示，本文计算单模态间的中心距差异（central

moment discrepancy, CMD)[5]作为共性学习损失,学习各模态的共性特征表示,并采用交叉注意力为单模态特征学习过程提供辅助信息.此外,本文以软注意力机制为基础,提出一种多模态注意力,增大强模态的表达,抑制弱模态对任务的影响.

1 相关工作

1.1 多模态情感分析

情感分析被广泛定义为对主观因素的计算研究,其中包括对人的意见、态度和情绪[6].多模态情感分析是利用多种形式的主观表达信息对情感进行分析判断的一种情感分析方法,当前的多模态情感分析方法主要针对多模态融合[7-8]和多模态表示进行研究[9-11].

多模态融合方面,Zadeh 等人[7]于 2017 年提出一种多模态张量融合方法对 3 种模态进行外积融合,得到的融合结果通过全连接深度网络进行最终的情感分析,该方法不仅能保存单个模态的内部信息,也能学习模态之间的互补信息;Tsai 等人[3]于 2019 年提出一种多模态转换器结构,该结构利用交叉模态注意力将文本、音频、图像 3 种模态进行交叉融合,使单个模态可以从其他的模态中获取信息.多模态表示方面,Hazarika 等人[9]于 2020 年提出一种模态共性和特性的表示方法,利用损失函数学习单模态表示之间的共性和特性,将这些模态间的共性表示和特性表示进行融合,减少了模态之间的信息冗余;Rahman 等人[10]于 2020 年提出一种多模态适应门的结构用于多模态模型的微调;Yu 等人[11]于 2021 年设计了一种自监督学习策略的标签生成模块,并以多任务学习的方式分别学习模态间的一致性表示和差异性表示.上述的多模态情感分析模型都是采用复杂的融合方式形成用于决策的多模态特征表示,没有考虑到模态间情感语义不一致问题对情感分析任务的影响.本文以交叉注意力和软注意力为基础,解决模态间情感语义不一致的问题,并提高强模态对任务的作用.

1.2 注意力机制

在现实生活中,人们会选择性地关注信息的重要部分,同时忽略其他部分,并以不同的重要程度对获取的信息进行处理[12].注意力模型最初用于机器翻译领域,目前在多个领域都有大量的应用[13].

空间注意力是一种自适应的空间区域选择机制,在图像、视频分类方面都起到了重要作用.Mnih 等人[14]提出循环注意力模型(recurrent attention model, RAM),采用循环神经网络和强化学习的方法使网络学习需要关注空间区域.Zhang 等人[15]提出了一种有效的全局注意力模块,该模块可以掌握全局和局部的观测信息,能更好地进行空间注意力学习.为了可视化图像字幕生成模型应该关注的位置和内容,Xu 等人[16]引入了软注意力机制为输入的信息分配权值,做加权平均处理.

时间注意力是一种动态的时序选择机制,它决定了视频中每一帧的重要性[17].Li 等人[18]提出了一种全局局部时间表示来利用自注意力机制获取视频序列中的多尺度时间线索.

Xiang 等人[19]于 2018 年以多种场景下的视频信息为基础,对 RGB 图、光流图、音频 3 种模态进行重要性分析,得出在不同场景下各模态体现的重要程度不同.Ghosal 等人[20]发现,在多模态任务的研究中,不是合并所有可用的模态都能有利于提高性能,也不是所有的模态都能在训练中发挥平等的作用.并针对这些问题提出基于多模态注意力框架的循环神经网络结构.

1.3 模态表示学习

在多模态情感分析领域中,模态表示学习可以被分为单模态表示学习和多模态表示学习.针对多模态表示学习,现有的研究往往会采用复杂的融合方法将多个模态融合为多模态表示,以获取模态间的交互关系[7-8].Ye 等人[21]采用对比学习的方式获取视觉模态和文本模态之间的交互语义信息,并在细粒度层面上融合 2 种模态信息.Sun 等人[22]提出一种多模态关系表达方法,利用各模态间的典型相关性学习模态间的共享信息,并利用共享信息学习具体的情感语义表达.针对单模态表示学习,Devlin 等人[23]提出预训练文本模型 BERT 用于提供更好的语言特征表示.Hazarika 等人[9]通过将每个模态映射到 2 个不同的子空间中学习各模态多种类型的特征表示.

2 本文模型

本文针对情感分析任务中不同模态间情感语义不一致会对多模态情感分析任务造成影响的问题,提出一种多模态学习模型,学习情感语义表达一致的模态特征表示,该模型的具体结构如图 1 所示.模型的具体学习过程如下:$\boldsymbol{X}_{\mathrm{T}}$, $\boldsymbol{X}_{\mathrm{A}}$, $\boldsymbol{X}_{\mathrm{V}}$分别代表 3 种模态的特征表示,模型利用两两模态的中心距差异值(CMD)作为损失函数对单模态网络的学习过程进行监督,学习到各模态的共性特征表示$\boldsymbol{C}_{\mathrm{T}}$, $\boldsymbol{C}_{\mathrm{A}}$, $\boldsymbol{C}_{\mathrm{V}}$;将

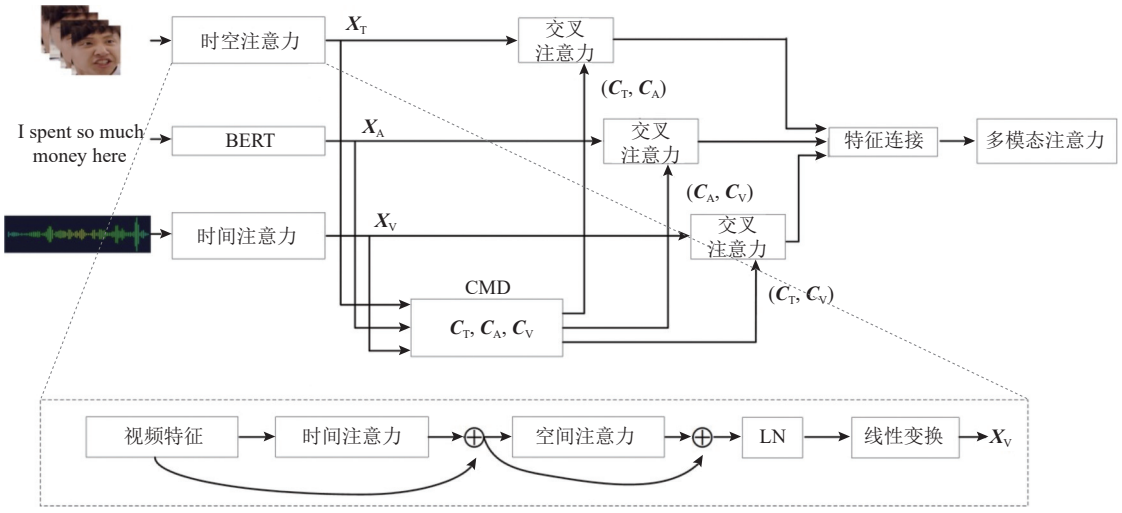


Fig. 1 Multimodal sentiment analysis model

图 1 多模态情感分析模型

原始的单模态特征表示$\boldsymbol{X}_{\mathrm{T}}$, $\boldsymbol{X}_{\mathrm{A}}$, $\boldsymbol{X}_{\mathrm{V}}$作为主要特征表示, $\boldsymbol{C}_{\mathrm{T}}$, $\boldsymbol{C}_{\mathrm{A}}$, $\boldsymbol{C}_{\mathrm{V}}$作为辅助特征表示输入到交叉注意力模块进行学习, 最终的输出由模态注意力进行加权连接.

2.1 特征提取

当前针对多模态情感分析的大多数研究仅单一地关注了模态中的时间信息或者空间信息, 忽略了视频模态中包含的时空信息. 为了获取空间中局部与全局的关系与时序特征中时间序列的长依赖关系, 本文采用时间注意力和时空注意力分别获取音频模态和视频模态的特征表示, 结构如图 1 所示.

令视频模态的特征表示为$\boldsymbol{X}_{\mathrm{V}} \in \mathbb{R}^{d\times F}$, 其中$d$是视频模态的特征维度, F是视频中的图像帧数. 为了获取空间中的局部信息, 将每帧图像的特征分解为N个均等长度且不重叠的块$\boldsymbol{p} \in \mathbb{R}^{d_i\times F}$, 使得$d = Nd_i$. 对每一个块有$\boldsymbol{Q}_p = \boldsymbol{W}_Q\boldsymbol{p}$, $\boldsymbol{K}_p = \boldsymbol{W}_K\boldsymbol{p}$, $\boldsymbol{V}_p = \boldsymbol{W}_V\boldsymbol{p}$, 则空间注意力可以表示为

$$\alpha_{\boldsymbol{p}}^{\mathrm{space}} = \frac{\boldsymbol{Q}_{\boldsymbol{p}}^{\mathrm{T}}}{\sqrt{d_i}}\left[\left\{\boldsymbol{K}_{\boldsymbol{p}^n}\right\}_{n=1,2,\cdots,N}\right], \tag{1}$$

$$\boldsymbol{s}_{\boldsymbol{p}}^{\mathrm{space}} = \sum_{n=1}^{N}\alpha_{\boldsymbol{p}}^{\mathrm{space}}\boldsymbol{V}_{\boldsymbol{p}^n}^{\mathrm{space}}. \tag{2}$$

时间注意力可以表示为

$$\alpha_{(\boldsymbol{p},t)}^{\mathrm{time}} = \frac{\boldsymbol{Q}_{(\boldsymbol{p},t)}^{\mathrm{T}}}{\sqrt{d_i}}\left[\left\{\boldsymbol{K}_{(\boldsymbol{p},t')}\right\}_{t'=1,2,\cdots,F}\right], \tag{3}$$

$$\boldsymbol{s}_{(\boldsymbol{p},t)}^{\mathrm{time}} = \sum_{t'=1}^{F}\alpha_{(\boldsymbol{p},t)}^{\mathrm{time}}\boldsymbol{V}_{(\boldsymbol{p},t')}^{\mathrm{time}}. \tag{4}$$

时空注意力即为上述 2 种注意力的结合.

2.2 中心距差异（CMD）

CMD 可以通过匹配 2 个模态特征表示的阶数矩差来度量 2 个模态的分布之间的差异, 其相比于 KL 散度包含了高阶矩信息, 相较于最大平均差异(maximum mean discrepancy, MMD)减少了计算量, 因此本文采用 CMD 学习模态间的共性特征表示. 2 种模态间的相似性越高, 对应的 CMD 就越小, 即

$$M_{\mathrm{CMD}}(\boldsymbol{r},\boldsymbol{q}) \to 0 \Rightarrow \boldsymbol{r} \xrightarrow{d} \boldsymbol{q}, \tag{5}$$

其中$M_{\mathrm{CMD}}(\boldsymbol{r},\boldsymbol{q})$代表$\boldsymbol{r},\boldsymbol{q}$这 2 个模态分布的中心距差异值, $\xrightarrow{d}$代表在分布上近似. 当$M_{\mathrm{CMD}}(\boldsymbol{r},\boldsymbol{q}) = 0$时, $\boldsymbol{r},\boldsymbol{q}$这 2 个模态分布相等, 即$\boldsymbol{r} = \boldsymbol{q}$.

本文研究中采用两两模态之间的 CMD 值作为损失, 用于在单模态表示学习的过程中学习到各模态的共性特征表示. 例如: 采用$M_{\mathrm{CMD}}(\mathrm{T},\mathrm{A})$作为文本模态和音频模态的中心距差异值, 则文本模态与音频模态的共性学习损失可以表示为$L_{(\mathrm{T},\mathrm{A})} = M_{\mathrm{CMD}}(\mathrm{T},\mathrm{A})$(中心距差异的计算满足交换性[5], 即$M_{\mathrm{CMD}}(\mathrm{T},\mathrm{A}) = M_{\mathrm{CMD}}(\mathrm{A},\mathrm{T})$). 在网络反向传播的过程中, 随着$L_{(\mathrm{T},\mathrm{A})}$的不断减小, 文本模态表示和音频模态表示之间的一致性会不断增加. 与上述相似, 在 3 种模态的表示学习过程中, $L_{(\mathrm{T},\mathrm{A})}$, $L_{(\mathrm{T},\mathrm{V})}$, $L_{(\mathrm{A},\mathrm{V})}$覆盖了所有模态之间的共性学习损失, 伴随着这 3 种损失的不断下降, 最终学到的 3 种单模态共性特征表示为$\boldsymbol{C}_{\mathrm{T}}$, $\boldsymbol{C}_{\mathrm{A}}$, $\boldsymbol{C}_{\mathrm{V}}$.

设有X和Y是 2 个独立且有界的随机样本, 在区间$[a,b]^N$上的概率分布为$\boldsymbol{r}$和$\boldsymbol{q}$. 则 2 个样本的中心矩差异为

$$M_{\mathrm{CMD}}(\boldsymbol{r},\boldsymbol{q}) = \frac{1}{|b-a|}\|\boldsymbol{E}(X)-\boldsymbol{E}(Y)\|_2 + \sum_{k=2}^{K}\frac{1}{|b-a|}\|c_k(X)-c_k(Y)\|_2, \tag{6}$$

其中$\boldsymbol{E}(X) = \dfrac{1}{X}\sum_{x\in X}x$是$X$的期望向量, $c_k(X) = E((x-$

$E(X))^k$)是所有第k阶样本的中心矩向量.

本文分别计算3种模态特征表示$\boldsymbol{X}_{\mathrm{T}}$, $\boldsymbol{X}_{\mathrm{A}}$, $\boldsymbol{X}_{\mathrm{V}}$两两之间的CMD值, 并将其作为损失函数的一部分, 学习3种模态的共性特征表示$\boldsymbol{C}_{\mathrm{T}}$, $\boldsymbol{C}_{\mathrm{A}}$, $\boldsymbol{C}_{\mathrm{V}}$. 共性学习损失函数的具体表示如下:

$$L=\frac{1}{3}\sum_{\substack{(m_1,m_2)\in\{(\mathrm{T,A}),\\(\mathrm{T,V}),(\mathrm{A,V})\}}} M_{\mathrm{CMD}}(m_1,m_2). \tag{7}$$

2.3 交叉注意力

在单模态表示学习的过程中, 为了在保留原始信息的情况下提高共性特征的表达, 同时增加模态间的动态交互, 本文采用交叉注意力使各模态的特征表示从其余模态的共性特征表示中获取辅助信息.

令文本模态、音频模态、视频模态分别为T, A, V. $\boldsymbol{X}_{\mathrm{T}}$, $\boldsymbol{X}_{\mathrm{A}}$, $\boldsymbol{X}_{\mathrm{V}}$分别为3种模态的特征表示. $\boldsymbol{C}_{X_{\mathrm{T}}}$, $\boldsymbol{C}_{X_{\mathrm{A}}}$, $\boldsymbol{C}_{X_{\mathrm{V}}}$分别为3种模态的共性特征表示. 将$\boldsymbol{X}_{\mathrm{T}}$作为主特征, $\boldsymbol{C}_{X_{\mathrm{A}}}$和$\boldsymbol{C}_{X_{\mathrm{V}}}$作为辅助特征($\boldsymbol{C}_{X_{\mathrm{A}}}$和$\boldsymbol{C}_{X_{\mathrm{V}}}$中既包含了各自模态的原始特征信息, 也包含了3种模态间的共性特征信息). 定义$\boldsymbol{Q}_{\mathrm{T}}=\boldsymbol{X}_{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{Q}_{\mathrm{T}}}$, 定义$\boldsymbol{K}_{\mathrm{A}}=\boldsymbol{C}_{X_{\mathrm{A}}}\boldsymbol{W}_{\boldsymbol{K}_{\mathrm{A}}}$, $\boldsymbol{V}_{\mathrm{A}}=\boldsymbol{C}_{X_{\mathrm{A}}}\boldsymbol{W}_{\boldsymbol{V}_{\mathrm{A}}}$, $\boldsymbol{K}_{\mathrm{V}}=\boldsymbol{C}_{X_{\mathrm{V}}}\boldsymbol{W}_{\boldsymbol{K}_{\mathrm{V}}}$, $\boldsymbol{V}_{\mathrm{V}}=\boldsymbol{C}_{X_{\mathrm{V}}}\boldsymbol{W}_{\boldsymbol{V}_{\mathrm{V}}}$, 其中$\boldsymbol{W}$为权值系数, $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$分别为特征经线性变换后的输出结果. 基于$\boldsymbol{Q}_{\mathrm{T}}$, $\boldsymbol{K}_{\mathrm{A}}$, $\boldsymbol{V}_{\mathrm{A}}$, $\boldsymbol{K}_{\mathrm{V}}$, $\boldsymbol{V}_{\mathrm{V}}$, 由音频模态A和视频模态V辅助文本模态T学习的过程可表示为:

$$y_{\mathrm{T}}=\frac{1}{2}(CrossM_{\mathrm{A}\to\mathrm{T}}+CrossM_{\mathrm{V}\to\mathrm{T}}), \tag{8}$$

$$CrossM_{\mathrm{A}\to\mathrm{T}}=softmax\left(\frac{\boldsymbol{Q}_{\mathrm{T}}\boldsymbol{K}_{\mathrm{A}}^{\mathrm{T}}}{\sqrt{dim_{X_{\mathrm{T}}}}}\right)\boldsymbol{V}_{\mathrm{A}}, \tag{9}$$

$$CrossM_{\mathrm{V}\to\mathrm{T}}=softmax\left(\frac{\boldsymbol{Q}_{\mathrm{T}}\boldsymbol{K}_{\mathrm{V}}^{\mathrm{T}}}{\sqrt{dim_{X_{\mathrm{T}}}}}\right)\boldsymbol{V}_{\mathrm{V}}, \tag{10}$$

其中y_{T}为文本模态特征提取网络的输出, 其余模态的单模态表示学习过程与文本模态相似. dim_X为输入特征的维度, 起到防止特征值过大的作用. $softmax$是归一化指数函数, 可表示为:

$$softmax(x_i)=\frac{\mathrm{e}^{x_i}}{\sum_{j=1}^{J}\mathrm{e}^{x_j}}. \tag{11}$$

2.4 模态注意力

本文将情感语义表达一致的单模态特征进行融合, 用于后期的任务决策. 为了增大强模态的表达, 抑制弱模态的影响, 将“软”注意力机制引入多模态融合的过程中, 提出一种模态注意力方法对各模态的特征表示进行加权连接. 具体结构如图2所示.

图2中, $\boldsymbol{T}^*$, $\boldsymbol{A}^*$, $\boldsymbol{V}^*$分别为经交叉注意力辅助后的3种模态特征表示. α_{T}, α_{A}, α_{V}分别为3种模态对任务的贡献度. 单模态对任务的贡献为

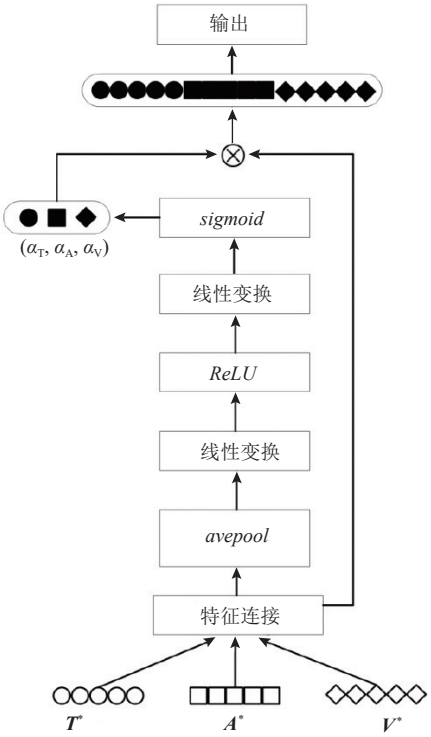


Fig. 2 Structure diagram of modal attention

图2 模态注意力结构图

$$\boldsymbol{h}=(avepool[\boldsymbol{T}^*,\boldsymbol{A}^*,\boldsymbol{V}^*])\boldsymbol{W}^{3\times3}, \tag{12}$$

$$[\alpha_{\mathrm{T}},\alpha_{\mathrm{A}},\alpha_{\mathrm{V}}]=sigmoid\left(\boldsymbol{h}\boldsymbol{W}^{3\times3}\right). \tag{13}$$

用于任务决策的多模态表示为

$$y=[\alpha_{\mathrm{T}}\boldsymbol{T}^*,\alpha_{\mathrm{A}}\boldsymbol{A}^*,\alpha_{\mathrm{V}}\boldsymbol{V}^*], \tag{14}$$

其中$sigmoid$为s型函数$\frac{1}{1+\mathrm{e}^{-x}}$, $[\cdot]$表示特征连接, $avepool$表示平均池化层.

3 实 验

3.1 数据集

针对多模态情感分析任务, 本文采用的数据集均包含3种模态, 分别为文本模态、音频模态、视频模态. 每个数据集的具体描述如下.

1)CMU-MOSI. 该数据集是由Zadeh等人[24]在2016年提出的第1个通过观点进行标注的情感分析数据集, 它包括了独白、演讲、电影等多种形式的观点数据, 共计93个视频、2 198个视频片段. 这些视频片段被人为手工标注至[–3, 3]之间, 作为情感评分, “–3”表示十分消极的情感, “3”表示十分积极的情感.

2)CMU-MOSEI. 该数据集是Zadeh等人[25]在2018年针对CMU-MOSI进行改进提出的情感分析数

据集. CMU-MOSEI 具有更多的样本数量，表达者和主题也具备了更大的多样性. 该数据集包含 23 453 个视频片段，分别来自于 5 000 个不同的视频.

3）CH-SIMS. 该数据集不仅包含了多模态表标签，而且包含了单模态标签，使研究人员可以研究单个模态之间的关系[2]. CH-SIMS 包含了 2 281 个视频片段，分别来自于 60 部不同种类的电影.

3.2 对比模型

为体现本文提出模型的有效性，将本文模型与现有的多模态情感分析模型进行对比. 对比模型如下.

1）张量融合网络（tensor fusion network, TFN）[7]. 该网络通过创建一个多维的张量，端到端学习模态内和模态间的动态信息.

2）低秩多模态融合（low-rank multimodal fusion, LMF）网络[8]. 该网络能减少由张量运算带来的复杂度，并获取模态内和模态间的信息.

3）多模态转换器 MulT[3]. 该方法利用交叉模态注意力机制获取多模态序列之间的远程交互，并使单模态可以从其他模态中获取信息.

4）模态的共性和特性表示（modality-invariant and specific representations, MISA）学习模型[9]. 该方法利用损失函数学习各个单模态表示之间的共性特征和特性特征，利用 Transformer 将学习到的特征进行融合.

5）BERT 多模态自适应门限（the multimodal adaptation gate for BERT, MAG-BERT）[10]. 该方法提出了一个基于 BERT 和 XLNet 微调模型的多模态自适应门，允许在微调的过程中接受多模态非语言数据.

6）自监督多任务学习（self-supervised multi-task learning, Self-MM）[11]. 该方法设计了一种单模态标签自动生成器，为每个模态提供单独的单模态训练任务.

7）基于多层感知器的情感分析模型 CubeMLP（an MLP-based model for multimodal sentiment analysis）[26]. 该方法通过 3 个独立的 MLP 单元输入并融合 3 种模态特征，并将混合的多模态特征以用于任务预测.

3.3 评估方法

本文采用相同的评估标准对模型在 CMU-MOSI 和 CMU-MOSEI 这 2 种数据集上得到的测试结果进行评估：2 分类（将数据标签分为积极和消极 2 类）、5 分类（5 类状态分别为消极、弱消极、中立、弱积极、积极）、7 分类（2 种数据集的标签都由人工标注为 [–3, 3] 的情感状态，其中 7 类状态由十分消极到十分积极逐渐变化）、$F1$ 分数（考虑了召回率和准确率的同等重要性）、MAE 和 Corr. 对于这些评估指标，除 MAE 外，值越高表示模型对于情感任务的分析性能越好.

3.4 超参数设置

为了获取合适的超参数，本文以验证集的结果为基础，采用网格搜索的方式对超参数进行选择. 最终实验的超参数设置如表 2 所示.

Table 2 Hyper-Parameter Settings in Dataset

表 2 数据集的超参数设置

超参数	MOSI	MOSEI	SIMS
batch size	32	64	64
learning rate	0.000 1	0.000 1	0.000 1
hidden size	256	256	256
depth	12	12	18
epoch	60	100	100
weight	0.5	0.5	0.8
early stop	20	20	20

注：“batch size”是单次训练的样本个数；“learning rate”是神经网络的学习率；“hidden size”是隐藏层的维度大小；“depth”是时空注意力和时间注意力中的学习层数；“epoch”是训练迭代次数；“weight”是共性特征表示学习的损失函数系数；“early stop”是为防止过拟合而设置的提前结束循环条件.

3.5 模型对比实验与结果分析

多模态情感分析的对比实验结果如表 3~5 所示.

Table 3 Comparison Results of Methods Based on CMU-MOSI Dataset

表 3 基于 CMU-MOSI 数据集的方法对比结果

模型	2 分类	5 分类	7 分类	$F1$	MAE	Corr
TFN[1]	0.808			0.807	0. 901	0.698
LMF[1]	0.825			0.824	0. 917	0.695
MulT[1]	0.834			0.828	0.871	0.698
MISA[1]	0.834			0.836	0.783	0.761
MAG-BERT[2]	0.843			0.843	0.731	0.789
Self-MM[2]	0.860			0.860	0.713	0.798
CubeMLP[3]	0.856		0.455	0.855	0.770	0.767
MAG-BERT*	0.835	0.501	0.438	0.836	0.749	0.783
Self-MM*	0.840	0.526	0.458	0.841	0.724	0.791
本文模型	0.866	0.532	0.474	0.867	0.709	0.796

注：上标为“1”的数据来源于文献 [9]，上标为“2”的数据来源于文献 [11]，上标为“3”的数据来源于文献 [26]，上标为“*”的数据结果是根据可获取的开源代码进行复现得出.

在实验中，本文首先与现有文献提供的数据结果进行对比，其次为了突显本文模型在分类任务上的有效性，增加 5 分类和 7 分类实验进行对比. 从实验结果中可以看出，本文提出的多模态情感分析方法在 CMU-MOSI, CMU-MOSEI, CH-SIMS 这 3 个数据

Table 4 Comparison Results of Models Based on CMU-MOSEI Dataset

表 4 基于 CMU-MOSEI 数据集的模型对比结果

模型	2 分类	5 分类	7 分类	*F*1	MAE	Corr
TFN[1]	0.825			0.821	0.593	0.700
LMF[1]	0.820			0.821	0.623	0.677
MulT[1]	0.825			0.823	0.580	0.703
MISA[2]	0.842			0.840	0.568	0.724
MAG-BERT[2]	0.852			0.851	0.539	0.753
Self-MM[2]	0.852			0.853	0.530	0.765
CubeMLP[3]	0.851		**0.549**	0.845	**0.529**	0.760
MAG-BERT*	0.851	0.545	0.527	0.850	0.543	0.760
Self-MM*	0.852	0.545	0.529	0.851	0.539	0.766
本文模型	**0.854**	**0.547**	0.531	**0.854**	0.537	**0.767**

注：黑体数字表示最优结果. 上标为“1”的数据来源于文献 [9]，上标为“2”的数据来源于文献 [11]，上标为“3”的数据来源于文献 [26]，上标为“*”的数据结果是根据可获取的开源代码进行复现得出.

Table 5 Comparison Results of Models Based on CH-SIMS Dataset

表 5 基于 CH-SIMS 数据集的模型对比结果

模型	多任务	2 分类	*F*1
TFN[2]	×	0.798	0.802
LMF[2]	×	0.794	0.787
Self-MM[2]	√	0.807	0.808
Self-MM*	√	0.792	0.789
本文模型	√	**0.812**	**0.825**

注：上标为“2”的数据来源于文献 [11]，上标为“*”的数据结果是根据可获取的开源代码进行复现得出.

集中的表现均优于其余模型.

为了更好地利用 CH-SIMS 数据集包含单模态标签的特点，本文模型采用多任务处理的方式在该数据集上进行评估，即增加单模态训练任务，利用单模态任务提升单模态的特征表达能力，并辅助多模态模型训练. 实验结果如表 5 所示.

3.6 有效性分析

为验证本文模型对解决模态间情感语义不一致问题的有效性，在保留 CH-SIMS 数据集中训练集和验证集的基础上，从测试集中选出 5 个包含模态情感语义不一致的样本进行测试，测试结果如表 6 所示. 其中，“实际值”为标签数据，预测值为模型测试结果；“N_{38}”表示 CH-SIMS 数据集中的第 38 个样本数据，以此类推.

从表 6 中可以看出，在第 38 和第 63 个样本中，文本模态和音频模态表现出消极的情感，与视频模态表现出的积极情感语义不一致. 在通过本文模型

Table 6 Analysis Results of Models Based on CH-SIMS Dataset

表 6 基于 CH-SIMS 数据集的模型分析结果

样本	实际值				预测值			
	T	A	V	M	T	A	V	M
N_{38}	–0.2	–0.6	0.4	–1.0	–0.4	–0.2	–0.2	–0.6
N_{63}	–0.6	–0.2	0.4	–0.6	–0.3	–0.2	–0.2	–0.5
N_{187}	–0.6	0.2	–1.0	–0.8	–0.5	–0.1	–0.3	–0.7
N_{732}	–0.8	–0.6	0.6	–0.6	–0.4	–0.2	–0.2	–0.6
$N_{2\,081}$	–0.2	0.6	–0.4	–1.0	–0.1	–0.2	–0.2	–0.2

注：“T”表示文本模态，“A”表示音频模态，“V”表示视频模态，“M”表示多模型.

进行学习后，3 种模态的情感语义趋于一致，解决了样本中模态情感语义不一致的问题. 在第 732 个样本中，视频模态表现出的情感与其余模态的情感不一致，且差距较大，而本文模型能有效利用其余模态对视频模态进行辅助，改变视频模态表达的情感语义，获取情感语义一致的多模态分析数据.

此外，本文采用余弦相似度对比模型训练前后的各模态差异. 余弦相似度的计算方法可由式(15)表示. 对比结果如表 7 所示.

$$sim(\boldsymbol{A},\boldsymbol{B})=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}},\tag{15}$$

其中 $\boldsymbol{A}$，$\boldsymbol{B}$ 是 2 个向量，$sim(\boldsymbol{A},\boldsymbol{B})$ 是 2 个向量的余弦相似度.

Table 7 Cosine Similarity Comparison of Modality Features Before and After Training

表 7 模态特征在训练前后的余弦相似度对比

数据集	训练前	训练后
MOSI	–0.011 8	0.031 4
MOSEI	–0.031 6	0.017 6
CH-SIMS	0.049 9	0.083 5

从表 7 中可以看出，本文模型能有效减小模态间的差异，并使不一致的情感语义趋于一致. 本文将情感语义一致的多模态特征数据用于多模态情感分析任务的后期决策.

3.7 方法对比

本文研究中采用两两模态之间的 CMD 值作为损失对单模态网络的学习过程进行监督，学习各模态的共性特征表示. 为了探究 CMD、MMD、KL 散度分别对获取共性特征表示模块的影响，本文以 MOSI

数据集为基础，将 MMD, KL 散度分别替换 CMD 方法进行计算，结果如表 8 所示.

Table 8 Comparison Results of Methods

表 8 方法对比结果

计算方法	2 分类	*F*1	MAE	Corr	训练时间/s
CMD	0.866	0.867	0.709	0.796	11
MMD	0.854	0.854	0.765	0.787	15
KL	0.846	0.847	0.770	0.764	9

由表 8 可以看出，CMD 由于减少了对核矩阵的计算，因此在运算时间上小于 MMD. 此外，使用 CMD 作为共性特征提取模块的任务表现明显高于其余 2 种方法 MMD 和 KL，证明了本文选择 CMD 方法的合理性.

3.8 消融分析

为进一步研究本文模型中单个模块对整体模型的影响，本文以 CMU-MOSI 数据集为基础，对模型中的模块进行消融分析，结果如表 9 所示.

Table 9 Results of Ablation Study

表 9 消融实验结果

模型	2 分类	*F*1	MAE	Corr
本文模型	0.867	0.867	0.709	0.796
(–) 时空注意力	0.836	0.835	0.763	0.786
(–) 交叉注意力	0.839	0.839	0.762	0.769
(–) 模态注意力	0.845	0.844	0.747	0.788

首先针对视频模态中时空注意力的使用进行消融分析. 消融实验中仅获取视频模态的时序信息而不获取空间信息. 从结果中看出，获取视频模态的时空信息更有利于任务的分析决策.

其次，针对交叉注意力模块进行消融分析. 消融实验中不利用共性特征进行辅助，而将学习到的共性特征作为输出. 从结果中可以看出，仅使用模态的共性特征表示会丢失大量的互补信息，导致实验结果下降.

最后，本文对提出的模态注意力模块进行消融分析. 消融实验中将各模态特征表示进行无权重连接，经线性变换后作为任务决策的多模态表示结果. 从表 9 中可以看出，增大强模态的权重并减小弱模态的影响有利于提升多模态任务的分析结果.

4 结 论

本文基于模态间情感语义表达不一致会对多模态情感分析任务造成影响的问题，提出一种多模态学习方法，学习情感语义表达一致的模态特征表示. 首先通过时间注意力和时空注意力分别提取音频模态和视频模态特征. 依据 CMD 计算量少、不需要计算核矩阵等优点，采用 CMD 计算模态之间的分布差距，学习各模态的共性特征表示. 利用交叉注意力使单个模态能从其余模态的共性特征表示中获取辅助信息，学习情感语义一致的单模态特征表示. 最后提出模态注意力对单模态特征表示进行加权连接，增大强模态的信息权重，减小弱模态对任务的影响. 本文将提出的多模态情感分析方法与现有的多模态情感分析模型在 CMU-MOSI, CMU-MOSEI, CH-SIMS 数据集上进行对比，结果显示本文提出的方法在所有评估标准中皆优于对比模型. 在未来的研究中，我们将进一步探索语义不一致问题对其他多模态任务的影响，并将该问题与不平衡的多模态学习问题进行结合.

作者贡献声明： 罗渊贻提出了算法思路，完成实验并撰写论文；吴锐提出了实验方案并修改论文；刘家锋、唐降龙提出指导意见并修改论文.

参 考 文 献

[1] Chen Long, Guan Ziyu, He Jinhong, et al. A survey on sentiment classification[J]. Journal of Computer Research and Development, 2017, 54(6): 1150−1170 (in Chinese)
(陈龙，管子玉，何金红，等. 情感分类研究进展[J]. 计算机研究与发展，2017, 54(6)：1150−1170)

[2] Yu Wenmeng, Xu Hua, Meng Fanyang, et al. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 3718−3727

[3] Tsai Y, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proc of the 57th Int Conf on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 6558−6569

[4] Bertasius G, Wang Heng, Torresani L. Is space-time attention all you need for video understanding[J]. arXiv preprint, arXiv: 2102.05095, 2021

[5] Zellinger W, Grubinger T, Lughofer E, et al. Central moment discrepancy (CMD) for domain-invariant representation learning[J]. arXiv preprint, arXiv: 1702.08811, 2017

[6] Sahay S, Kumar S H, Xia Rui, et al. Multimodal relational tensor network for sentiment and emotion classification[C]//Proc of the 1st Int Conf on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 20−27

[7] Zadeh A, Chen Minghai, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proc of the Int Conf on Empirical

Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 1103−1114

[8] Liu Zhun, Shen Ying, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[C]//Proc of the Int Conf on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2247−2256

[9] Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis[J]. arXiv preprint, arXiv: 2005.03545, 2020

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

[11] Yu Wenmeng, Xu Hua, Yuan Ziqi, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[C]//Proc of the Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 10790−10797

[12] Cheng Yan, Yao Leibo, Zhang Guanghe, et al. Text sentiment orientation analysis of multi-channels CNN and BiGRU based on attention mechanism[J]. Journal of Computer Research and Development, 2020, 57(12): 2583−2595 (in Chinese)
(程艳, 尧磊波, 张光河, 等. 基于注意力机制的多通道 CNN 和 BiGRU 的文本情感倾向性分析[J]. 计算机研究与发展, 2020, 57(12): 2583−2595)

[13] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint, arXiv: 1409.0473, 2014

[14] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C]//Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: NeurIPS, 2014: 2204–2212

[15] Zhang Zhizheng, Lan Cuiling, Zeng Wenjun, et al. Relation-aware global attention for person re-identification[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3183–3192

[16] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proc of the Int Conf on Machine Learning. New York: ACM, 2015: 2048−2057

[17] Guo Menghao, Xu Tianxing, Liu Jiangjiang, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331−368

[18] Li Jianing, Wang Jingdong, Gao Wen, et al. Global–local temporal representations for video person re-identification[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 3957–3966

[19] Xiang Long, Gan Chuang, Melo G, et al. Multimodal keyless attention fusion for video classification[C]//Proc of the 32nd Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 7202−7209

[20] Ghosal D, Akhtar M S, Chauhan D, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 3454−3466

[21] Ye Junjie, Zhou Jie, Tian Junfeng, et al. Sentiment-aware multimodal pre-training for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2022, 258: 110021

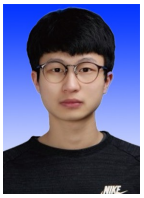
[22] Sun Zhongkai, Sarma P, Sethares W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]//Proc of the Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 8992−8999

[23] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the Conf on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4171–4186

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

[25] Zadeh A, Liang P P, Vanbriesen J, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proc of the 56th Conf on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2236−2246

[26] Sun Hao, Wang Hongyi, Liu Jiaqing, et al. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 3722−3729



Luo Yuanyi, born in 1996. PhD candidate. His main research interest includes multimodal learning.
罗渊贻, 1996 年生. 博士研究生. 主要研究方向为多模态学习.



Wu Rui, born in 1976. PhD, associate professor, PhD supervisor. His main research interests include pattern recognition and multimodal learning.
吴　锐, 1976 年生. 博士, 副教授, 博士生导师. 主要研究方向为模式识别、多模态学习.



Liu Jiafeng, born in 1968. PhD, associate professor. His main research interests include pattern recognition and machine learning.
刘家锋, 1968 年生. 博士, 副教授. 主要研究方向为模式识别、机器学习.



Tang Xianglong, born in 1960. PhD, professor, PhD supervisor. His main research interests include pattern recognition and computer vision.
唐降龙, 1960 年生. 博士, 教授, 博士生导师. 主要研究方向为模式识别、计算机视觉.