文章编号：2096-1618(2022)06-0656-06

基于多模态融合的视频情感分析技术

陈诗汉， 马洪江， 王 婷， 何松泽
(成都信息工程大学计算机学院，四川 成都 610200)

摘要：介绍一种视频多模态情感识别方法。一个视频通常通过文本、声音和视觉图像等多模态信息来表达同一种情感主题，而如何将同一个视频中不同异构数据之间的信息融合并最大程度地利用是目前需要重点攻克的难题。通过互信息最大化的方法，高效融合视频中的文本、声音与视觉图像等多模态异构数据，尽可能多地消除模态之间的差异，最终实现对视频的情感进行识别分析。在公开的 MOSEI 多模态数据集上进行实验，实验结果显示 MAE 值达55.4。相比之前的一些模型，本模型效果更优，且实验模型构造不繁琐，为后面相关的研究打下良好的基础。

关 键 词：多模态融合；视频情感分析；互信息最大化
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.06.007

0 引言

近年来随着社交媒体的快速发展以及智能手机的普及，多模态数据呈爆炸式增长，如图像、视频等。多模态数据是用户交流和记录生活的媒介，通常蕴涵着丰富的个人情感。从多模态数据中挖掘和理解情感信息，即多模态情感分析(multimodal sentiment analysis, MSA)，已经成为一个热门的研究课题。相较于传统的文本情感信息提取，对于视频这类的多模态数据提取会存在很多困难，因为其包含了语音、文本以及图像信息。而且传统的基于单模态情感分析的机器学习方法在多模态情感分析这类任务上存在较大的局限性[1]。

鉴于人可以用不同的方式表达情感，包括使用不同的声调或面部表情，对于这些多模态数据，同一数据段中的不同模态会相互补充[2]，为语义和情感消歧提供额外帮助。因此可以使用多模态融合相关技术来识别人类的情感[3]。多模态融合技术是一种从海量多模态数据中提取整合信息并可用于提高信息处理效率的技术[4]，现已被广泛用于处理结构化数据和文本数据[5]。目前该领域的大部分工作都集中在早期或晚期融合上。早期的融合模型采用简单的网络架构，Zadeh 等[6]提出了一个张量融合网络，在更深层融合了不同的模态表征。薛其威等[7]通过多模态特征融合对无人驾驶系统车辆进行检测，在 KITTI 数据集上其平均检测精度为84.71%。另外，Sun 等[8]优化了模态表征之间的相关性以进行融合，然后将融合结果传递给下游任务。

受深度学习的影响，各类相关研究层出不穷，其中注意力机制获得广泛关注，LSTM(long short-term memory)被用于随时间捕获模态之间的交互。颜增显等[9]利用多模态通道注意力网络来融合不同模态的特征进行人脸反欺骗算法研究，在 CASIA-SURF 数据集上获得良好的效果。王旭阳等[10]利用注意力机制与时域卷积网络建立多模态融合的模型，在 CMU-MOSI 数据集上相较于基线有了较大的提升。Tsai 等[11]提出一种可以动态调整模态之间的权重，为多模态融合提供可解释性的方案。受模态分离领域进步的推动，Hazarika 等[12]将模态特征投影到专有和公共特征空间中，以捕捉不同模态的独有和共享特征以方便后期进行融合。虽然这些研究中能达到的效果比较有限，但也为后续相关研究做好了相应的铺垫。Makiuchi 等[13]提出了一种基于 Transformer 的模型将语音和文本数据进行融合，在 IEMOCAP 数据集上得到73.0%的准确率。Byun 等[14]也提出了一种利用深度学习融合语音和文本数据进行情感识别的模型，在自行构建的韩语数据集上达到了95.97%的准确率。还有黄欢等[15]设计了一个 AV-MSA 模型，利用交叉投票机制将视觉与音频信息融合进行情感分析，在 IEMOCAP 和 WB-AV 数据集上取得了较好的效果，这些研究表明情感识别任务可以从多模态中受益。

在 MSA 任务中进行信息抽取以及信息融合的时候可能会丢失实际信息并额外引入每种模态携带的噪声。为减少这个问题带来的影响，一种互信息(mutual information, MI)方法被用于评估成对的多维变量(即各个模态)之间的依赖关系，并且可有效去除与下游任务无关的冗余信息[16]。由于互信息在处理时，会存

收稿日期：2022-07-19
基金项目：四川省科技厅重点研发资助项目(2021YFG0031、2022YFG0375)；四川省科技服务业示范资助项目(2021GFW130)

在信息丢失的问题[17]。本文基本互信息方法提出了一种多模态融合最大化模型(multi-modal fusion max, MMFM),其核心是在多模态融合中分层最大化互信息。

本文提出一种基于多模态融合的分层 MI 最大化模型,用于多模态情感分析。其中多模态融合最大化发生在输入和融合模块,可以减少有价值任务相关信息的丢失。在公开的情感数据集上进行的实验,获得较好的效果。

1　方法

1.1　概述

在多模态情感分析任务中,模型的输入是从视频片段中提取的单模态原始序列 X_m,其中 m 表示向量维数。文中,$m \in \{t, v, a\}$,其中 t,v,a 分别表示 3 种不同类型的模态——文本、视觉和声音。目标是从这些输入向量中提取和整合关于任务相关的情感信息,形成统一的表示,并将其用于对反映情感强度的真值 y 进行准确预测。

1.2　整体架构

整体框架结构如图 1 所示,输入的信息包括视频、文本和语音 3 种。首先,模型使用特征提取器和编码器分别将 3 种原始输入处理为数字序列向量 X_v,X_a,X_t。然后,编码后的数据主要经过融合网络和 MI 最大化两部分进行处理,分别对应着图 1 中的实线和虚线标记。其中,在融合部分融合网络将不同模态信息两两交互,将单模表示转换为融合结果 K,再通过回归多层感知器(multi-layer perception, MLP)进行最终的预测。在互信息部分,MI 最大化是为了估计和提升输入层和融合层的 MI 下界。这两个部分同时工作用于产生后续识别任务以及互信息相关的损失,通过模型学习将任务相关信息融入融合结果,并提高主任务中预测的准确性。

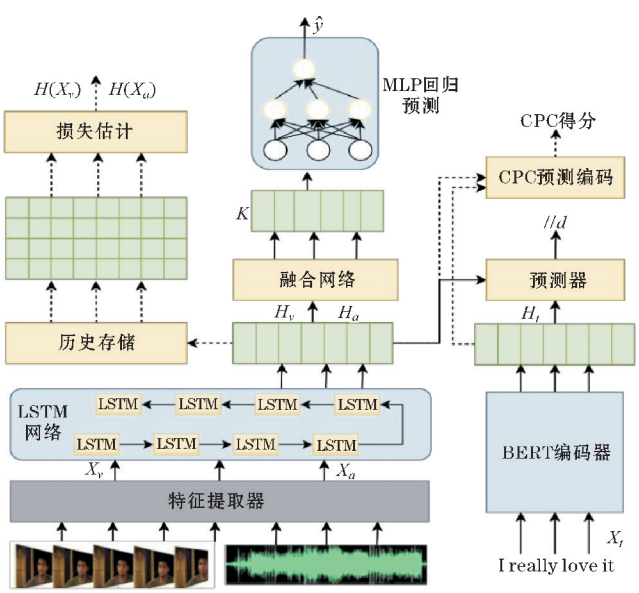

图 1　模型总体结构

1.3　模态编码

模态编码负责将多模态顺序输入 X_m 编码为单位长度表示为 H_m 具体来说,对于文本信息,使用 BERT[18](bidirectional encoder representation from transformers)对输入句子进行编码,并从最后一层的输出中提取头部嵌入作为 H_t。对于视觉和声学的内容,采用两种特定于模态的单向 LSTM[19]捕捉这些模态的时间特征。

1.4　模态间 MI 最大化

互信息是信息论中的一个概念,用于估计变量对之间的关系[20],定义为

$$I(X;Y)=\sum_{x,y}p(x,y)\left[\log\frac{p(x,y)}{p(x)p(y)}\right] \tag{1}$$

其中 x 与 y 为不同的随机变量。

Alemi 等[21]首先将与 MI 相关的优化结合到深度学习模型中。另外在其他研究中也有证明 MI 最大化的优势[22]。然而,由于在高维空间中直接进行 MI 几乎是不可能的,所以很多工作都是直接优化 MI 的下界。文中,是在输入级别和融合级别应用 MI 下界,并根据要估计的项的数据特征和数学属性制定这些界限的估计方法。

MI 可以评估视频中不同模态间的依赖程度,通过将 MI 最大化可以实现多模态间更好的融合。对于视频 V,将来自单个视频剪辑的模态表示对标记为 X 和 Y(它们之间通常存在相关性),在先验分布已知时,可以将 X 和 Y 的先验分布化为 $P(X)=\int_V P(X,Y|V)P(V)$,$P(Y)=\int_V P(Y|V)P(V)$,联合分布为 $P(X,Y)=\int_V P(X,Y|V)P(V)$。因存在相关性,可以利用 MI 过滤掉与任务无关的噪声来提高性能。基于以上分析,为实现多模态更大程度的融合并且保持模态内容不变,本文利用一个易于处理的 MI 下限,而不是直接计算 MI,并参照 Baber 等[23]采用的较为准确且直接的 MI 下限,其近似于真值条件分布 $p(y|x)$,如式(2)所示。

$$\begin{aligned} I(X;Y)=&\sum_{x,y}p(x,y)\log\left[\frac{q(y|x)}{p(y)}\right]+\\ &\sum_{x,y}p(y)\left[KL(p(y|x)\,||\,q(y|x)\right]\\ \geqslant&\sum_{x,y}p(x,y)\left[\log q(y|x)\right]+H(Y)\triangleq I_B \end{aligned} \tag{2}$$

其中 $q(y|x)$ 是预测的概率分布，KL 是用于度量两个概率分布相似度的指标，$H(Y)$ 是 Y 的微分熵，I_B 为 Baber 等使用的 MI 下界。当 $q(y|x) = p(y|x)$ 时，界值和真值之间没有差距。在每一对模态 (X,Y) 中，其中一个模态视为 X，则另外一个模态视为 Y。然后训练一个预测器 $q(y|x)$ 来逼近 $p(y|x)$。本文在实验过程中优化了不同模态对的边界—文本与视觉、文本与声学、视觉与声学。另外，在消融研究部分检查了设计的有效性。将 $q(y|x)$ 公式化为多元高斯分布 $q_\theta(y|x)=N(y|\mu_{\theta_1}(x),\sigma^2_{\theta_2}(x)I)$，两个神经网络由 θ_1 和 θ_2 参数化为分别预测均值和方差。损失函数为：

$$\mathcal{L}_{lld}=-\frac{1}{n}\sum_{tv,ta,va}\sum_{i=0}^{n}\log q(y_i|x_i) \tag{3}$$

其中 n 是训练中的批量大小，tv，ta，va 表示 3 个预测变量的可能性之和。

本文采用情感极性（非负/负）作为分类标准，它是数据集中的一个自然属性，可以平衡估计精度和计算成本。对于熵项 H(Y)，使用高斯混合模型（Gaussian mixed model，GMM）来求解计算，这是一种常用的未知分布近似方法。GMM 为不同的属性类别建立了多个高斯分布。多元正态分布的熵为：

$$H=\frac{1}{2}\log((2\pi e^k\det(\boldsymbol{\Sigma}))=\frac{\log((\det 2\pi e\,\boldsymbol{\Sigma}))}{2} \tag{4}$$

式中 k 是 GMM 中向量的维数，$\det(\boldsymbol{\Sigma})$ 是协方差矩阵 $\boldsymbol{\Sigma}$ 的行列式。基于数据集中两个极性类别的频率几乎相等，本文采用来自 Huber 等[24]使用的 GMM 熵的下界和上界，公式如下：

$$\sum_c w_c h_c \leqslant H(Y) \leqslant \sum_c w_c(-\log w_c + h_c) \tag{5}$$

其中 h_c 是 c 类的子分布的熵，w_c 为 c 类子分布的先验概率。取下界作为近似值，得到 MI 下界的熵项：

$$H(Y)=\frac{1}{4}[\log((\det\boldsymbol{\Sigma}_2)\det\boldsymbol{\Sigma}_2)] \tag{6}$$

另外，在训练时，根据统计理论，应该增加批量大小以减少估计误差，可以通过包含最近历史的数据来间接扩大采样批次。在实验过程中将这些数据存储在历史数据存储器中，MI 下限最大化的损失函数由式(7)给出：

$$\mathcal{L}_B=-I_B^{t,v}-I_B^{t,a}-I_B^{t,v} \tag{7}$$

1.5 融合层面的 MI 最大化

为捕获模态之间的模态不变线索，在融合结果和输入模态之间重复 MI 最大化。目标是产生融合结果 $K=F(X_t,X_v,X_a)$ 的融合网络 F。由于已经有了从 X_m 到 K 的生成路径，考虑一条相反的路径，即从 K 构造 X_m，$m\in\{t,v,a\}$。可以使用分数函数作用于归一化的预测和真值向量来衡量它们的相关性：

$$s(h_m,K)=\exp(\overline{h_m}(\overline{G_\varphi(K)})^{\mathrm{T}}) \tag{8}$$

其中 G_φ 是参数 φ 的神经网络，它从 K 生成 H_m 的预测，通过将同一批次中该模态的所有其他表示 $\widetilde{H}_m^i=H_m\backslash\{h_m^i\}$ 视为负样本，将这个分数函数合并到噪声对比估计框架[25]中，即

$$\mathcal{L}_N(K,H_m)=-\sum_H\log\frac{s(K,h_m^i)}{\sum_{h_c^j\in H_m}s(K,h_m^j)} \tag{9}$$

等式(9)实际上视为二分类交叉熵损失，H 是一组样本，公式中分数上下两部分可以视为正负样本对，当正样本对之间的互信息更大，负样本对之间的互信息更小时，符合互信息最大化要求，因此通过优化该损失，可以让互信息最大化。由于对比预测编码（contrastive predictive coding，CPC）可以学习更多的全局结构，在模型中，融合结果 K 反向预测跨模态的表示，以便可以将更多模态固有信息传递给 K。此外，通过将每个模态的预测对齐，使模型能够决定它应该从每种模态中接收到多少信息。损失函数为

$$\mathcal{L}_{\mathrm{CPC}}=\mathcal{L}_N^a+\mathcal{L}_N^v+\mathcal{L}_N^t \tag{10}$$

1.6 训练

训练过程包括两个阶段：在第一阶段，近似p(y|x)与q(y|x)通过最小化多模态预测变量的负对数似然。在第二阶段，将之前的 MI 下界作为辅助损失添加到主要损失中。在获得最终预测 $\hat{y}$ 及真值 y 后得到任务损失：

$$\mathcal{L}_{\mathrm{task}}=\mathrm{MAE}(\hat{y},y) \tag{11}$$

其中 MAE（mean absolute error）代表平均绝对误差损失。最后来计算所有这些损失的加权和以获得该阶段的主要损失：

$$\mathcal{L}_{\mathrm{main}}=\mathcal{L}_{\mathrm{task}}+\alpha\mathcal{L}_{\mathrm{cpc}}+\beta\mathcal{L}_B \tag{12}$$

其中 α、β 是控制 MI 最大化影响的超参数。

2 实验

2.1 数据集

采用数据集为关于多模态情感分析研究的公开数据集，即 CMU-MOSEI[26]，它包含来自 YouTube 的 23454 个电影视频剪辑。

2.2 基本设置与指标

本文分别采用 P2FA[27]和 COVAREP[28]工具包对于图像和音频内容进行特征提取。而对于文本内容，使用预训练好的 BERT 模型来获得词向量，最后在 GPU 上训练模型。评测指标如下：平均绝对误差（MAE），它是预测值和真值之间的平均绝对差值，衡

量预测偏斜程度的皮尔逊相关性(pearson correlation, Corr),七分类准确度(seven-classclassification accuracy, Acc-7),二分类准确度(binary classification accuracy,Acc-2)和 $F1$ 分数。

2.3　模型比较

为了解本文模型的相对性能,将模型与许多具有较好效果的基线进行比较,如 TFN[14]、LMF[29]、MFM[16]、MULT[11]、ICCN[30]和 MISA[13]。

2.4　结果

实验结果见表 1 所示,对于 Acc-2 和 $F1$ 值有两组评估结果,左边值为积极情绪结果,右边值为消极情绪结果,可以发现 MMFM 与许多基线方法相比具有更优的结果。具体来说,本文模型在 CMU-MOSEI 上的 Acc-7、Acc2、$F1$ 得分都优于其他模型。对于其他指标,MMFM 的性能也非常好。这些结果初步证明了本文的方法在多模态情感分析任务中的有效性。

表 1　CMU-MOSEI 数据集上的运行结果

模型	MAE	Corr	Acc-7	Acc-2	$F1$
TFN	0.593	0.700	50.2	–/82.5	–/82.1
LMF	0.623	0.677			
MFM	0.568	0.717	51.3	–/84.4	–/84.3
ICCN	0.5653	0.713	51.6	– /84.2	– /84.2
MULT	0.580	0.703		– /82.5	– /82.3
MISA	0.568	0.724		82.59/84.23	82.67/83.97
Ours	0.554	0.752	52.9	83.3/84.9	83.5/84.7
MAG-BERT	0.539	0.765		83.8/85.2	83.7/85.1

2.5　消融研究

为体现模型中提出的损失函数和相应估计方法的优势,本文在 CMU-MOSEI 上进行了一系列消融实验,表 2 为不同消融设置下的结果。首先,消除了一个或几个 MI 损失项,包括模态间的 MI 下限(I_B)和 CPC 损失。从表 2 中可以注意到去除部分 MI 损失后明显的性能下降,它显示了多模态融合最大化模型的效果。此外,通过将多模态 MI 中的当前优化目标对替换为单个对或其他对组合,无法获得更好的结果,也验证设计的合理性。然后测试熵估计,当停用历史记忆并仅使用当前批次评估中的 μ 和 Σ 时,出现“NaN”值,表示训练过程崩溃。因此,基于历史的估计具有保证训练稳定性的优点。最后,将 GMM 替换为统一的高斯分布,其中 μ 和 Σ 在所有样本上进行估计,不管它们的极性类别如何,结果发现所有指标都有明显下降,这意味着基于自然分类的 GMM 可以更准确地估计熵项。

表 2　模型消融研究结果

描述	MAE	Corr	Acc-7	Acc-2	$F1$
Ours	0.554	0.752	52.9	83.3/84.9	83.50/84.70
模态内融合					
$I_B^{t,v}$	0.571	0.728	51.01	81.01/84.73	81.56/84.75
$I_B^{t,a}$	0.558	0.739	52.52	79.42/83.49	80.12/83.57
$I_B^{v,a}$	0.556	0.748	52.01	82.8/84.45	82.55/84.49
$I_B^{t,v}+I_B^{t,a}$	0.560	0.752	52.75	75.63/81.83	74.48/81.59
$I_B^{v,a}+I_B^{t,v}$	0.559	0.749	52.71	82.19/84.98	82.56/84.90
$I_B^{v,a}+I_B^{t,a}$	0.606	0.708	49.09	81.88/83.98	82.06/83.74
none	0.574	0.729	51.8	74.48/81.59	75.63/81.83
L_{CPC}损失					
有/无 $L_N^{k,t}$	0.570	0.728	52.02	74.97/81.18	76.10/81.45
有/无 $L_N^{k,v}$	0.562	0.727	51.89	82.12/83.87	82.42/83.78
有/无 $L_N^{k,a}$	0.576	0.733	51.38	69.49/77.43	68.08/77.05
有/无 $L_N^{k,t}$,$L_N^{k,v}$,$L_N^{k,a}$	0.574	0.729	51.83	72.29/82.22	78.19/82.38
损失评估					
有/无历史数据	NaN	NaN	NaN	NaN/NaN	NaN/NaN
有/无 GMM	0.578	0.733	51.01	81.41/83.41	81.12/82.24

3 结论

从模型在数据集上的表现来看,本文提出的多模态最大化融合框架在针对多模态情感识别的问题上取得一定的效果。且进一步的消融研究结果验证了模型的有效性。在未来,将多模态应用于情感分析会有较好的发展潜力以及较高的应用价值,相信这项工作可以更多激发多模态情感分析的创造力。

参考文献:

[1] 奚晨. 基于表情、语音和文本的多模态情感分析[D]. 南京:南京邮电大学,2021.

[2] 王蝶. 基于注意力机制的多模态融合技术研究[D]. 南京:南京师范大学,2021.

[3] 冯亚琴,沈凌洁,胡婷婷,等. 利用语音与文本特征融合改善语音情感识别[J]. 数据采集与处理,2019,34(4):625-631.

[4] 秦放,曾维佳,罗佳伟,等. 基于深度学习的多模态融合图像识别研究[J]. 信息技术,2022(4):29-34.

[5] 牟智佳,符雅茹. 多模态学习分析研究综述[J]. 现代教育技术,2021,31(6):23-31.

[6] Zadeh A,Chen M,Poria S,et al. Tensor FusionNetwork for Multimodal Sentiment Analysis[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017:1103-1114.

[7] 薛其威,伍锡如. 基于多模态特征融合的无人驾驶系统车辆检测[J]. 广西师范大学学报(自然科学版),2022,40(2):37-48.

[8] Sun Z,Sarma P,Sethares W,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. 2020,34(5):8992-8999.

[9] 颜增显,孔超,欧卫华. 基于多模态融合的人脸反欺骗算法研究[J]. 计算机技术与发展,2022,32(4):63-68.

[10] 王旭阳,董帅,石杰. 复合层次融合的多模态情感分析[J/OL]. http://kns.cnki.net/kcms/detail/11.5602.TP.20220331.1739.003.html,2022(8):31.

[11] Tsai Y H,Bai S,Kolter J Z,et al. Multimodal Transformer for Unaligned Multimodal Language Sequences[C]. Proceedings of the conference. Association for Computational Linguistics,2019,2019:6558-6569.

[12] Hazarika D,Zimmermann R,Poria S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis[C]. Proceedings of the 28th ACM international conference on multimedia,2020:1122-1131.

[13] Makiuchi M R,Uto K,Shinoda K. Multimodal emotion recognition with high-level speech and text features[C]. 2021 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU). IEEE,2021:350-357.

[14] Byun S W,Kim J H,Lee S P. Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding[J]. Applied Sciences,2021,11(17):7967.

[15] 黄欢,孙力娟,曹莹,等. 基于注意力的短视频多模态情感分析[J]. 图学学报,2021,42(1):8-14.

[16] Poole B,Ozair S,Van Den Oord A,et al. On variational bounds of mutual information[C]. International Conference on Machine Learning. PMLR,2019:5171-5180.

[17] Belghazi M I,Baratin A,Rajeshwar S,et al. Mutual information neural estimation[C]. International conference on machine learning. PMLR,2018:531-540.

[18] Devlin J,Chang M W,Lee K,et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Proceedings of NAACL-HLT,2019:4171-4186.

[19] Hochreiter S,Schmidhuber J. Long short-term memory[J]. Neural computation,1997,9(8):1735-1780.

[20] Tishby N,Zaslavsky N. Deep learning and the information bottleneck principle[C]. 2015 ieee information theory workshop. IEEE,2015:1-5.

[21] Alemi A A,Fischer I,Dillon J V,et al. Deep Variational Information Bottleneck[J]. arXiv e-prints,2016:arXiv:1612.00410,2016.

[22] Bachman P,Hjelm R D,Buchwalter W. Learning representations by maximizing mutual information across views[C]. Proceedings of the 33rd International Conference on Neural Information Pro-

cessing Systems,2019:15535-15545.

[23] Barber D, Agakov F. The IM algorithm: a variational approach to Information Maximization[C]. Proceedings of the 16th International Conference on Neural Information Processing Systems, 2003: 201-208.

[24] Huber M F, Bailey T, Durrant-Whyte H, et al. On entropy approximation for Gaussian mixture random vectors[C]. 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems. IEEE, 2008: 181-188.

[25] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models[C]. Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 297-304.

[26] Zadeh A A B, 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, 2018: 2236-2246.

[27] Yuan J, Liberman M. Speaker identification on the SCOTUS corpus[J]. The Journal of the Acoustical Society of America, 2008, 123(5): 3878-3878.

[28] Degottex G, Kane J, Drugman T, et al. COVAREP—A collaborative voice analysis repository for speech technologies[C]. 2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014: 960-964.

[29] Yu W, Xu H, Yuan Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35 (12): 10790 - 10797.

[30] 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 (Volume 1: Long Papers), 2018: 2247-2256.

Video Sentiment Analysis Technology based on Multimodal Fusion

CHEN Shihan, MA Hongjiang, WANG Ting, HE Songze

(College of Computer, Chengdu University of Information Technology, Chengdu 610200, China)

Abstract: A method for multimodal sentiment recognition in video is introduced in this paper. A video usually expresses the same sentiment theme through multimodal information such as text, sound, and visual images, andfusingthe information between different modalities and make full use of them is the current key problems that need to be overcome. This paper uses the method of maximizing mutual information to efficiently fuse multimodal heterogeneous data such as text, sound and visual images in videos to eliminate as many differences between modalities as possible, and finally realize the recognition and analysis of video sentiment. Experiments are carried out on the public MOSEI multimodal dataset, and the results show that the MAE value reaches 55.4. Compared with conventional models, the effect of this model is better, and the construction of the experimental model is not cumbersome, which can provide reference for related research.

Keywords: multimodal fusion; video sentiment analysis; mutual information maximization