天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于两阶段缺失模态恢复的多模态情感分析方法
作者：王媛，邓振宇，王佳鑫，张帅，赵婷婷，于琦
DOI：10.13364/j.issn.1672-6510.20230188
收稿日期：2023-10-15
网络首发日期：2024-06-21
引用格式：王媛，邓振宇，王佳鑫，张帅，赵婷婷，于琦．基于两阶段缺失模态恢复的多模态情感分析方法[J/OL]．天津科技大学学报．
https://doi.org/10.13364/j.issn.1672-6510.20230188

DOI:10.13364/j.issn.1672-6510.20230188

基于两阶段缺失模态恢复的多模态情感分析方法

王 媛[1]，邓振宇[1]，王佳鑫[1]，张 帅[1]，赵婷婷[1]，于 琦[2]
（1. 天津科技大学人工智能学院，天津 300457; 2. 山西医科大学管理学院，太原 030607）

摘 要：在多模态情感分析任务中，模态缺失情况下存在整体多模态与单模态情感表达不一致、情感会因模态缺失而发生变化的情况。现有方法忽略缺失模态潜在的特异性语义内涵和情感表达，导致情感分析性能下降。为解决这个问题，本文提出基于两阶段缺失模态恢复的多模态情感分析模型。首先恢复缺失模态特征，通过不变模态翻译模块恢复缺失模态与文本模态的同构信息，接着通过相似模态翻译模块在视觉词典中恢复缺失模态的异构信息，然后将已知模态和恢复出的缺失模态进行融合，实现一致性的情感分析。在 CMU-MOSI 和 IEMOCAP 上的实验表明，该模型能够有效检测和恢复缺失模态语义特征，缓解多模态与单模态情感表达不一致的问题，并具有显著的性能优势。

关键词：多模态；情感分析；模态缺失
中图分类号：TP391.4 **文献标志码：**A **文章编号：**1672-6510（0000）00-0000-00

Multimodal Sentiment Analysis Approach Based on Two-Stage Missing Modality Recovery

WANG Yuan[1]，DENG Zhenyu[1]，WANG Jiaxin[1]，ZHANG Shuai[1]，ZHAO Tingting[1]，YU Qi[2]
(1. College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China；
2. College of Management, Shanxi Medical University, Taiyuan 030607, China)

Abstract: In the task of multimodal sentiment analysis, there are cases where the overall multimodal and unimodal sentiment expressions are inconsistent in the case of missing modalities, and the sentiment will change due to missing modalities. Existing methods ignore the potential specific semantic connotations and sentiment expressions of the missing modality, resulting in degradation of sentiment analysis performance. To solve this problem, in this article we propose a multimodal sentiment analysis model based on two-stage missing modality recovery. In our proposed model, the missing modal features are first recovered, recover the isomorphic information of the missing modality with the textual modality through the invariant modal translation module, followed by recovering the heterogeneous information of the missing modality in the visual dictionary through the similar modal translation module, and then the known modalities and the recovered missing modalities are fused to achieve consistent sentiment analysis. Experiments on CMU-MOSI and IEMOCAP showed that the model could effectively detect and recover missing modality semantic features, alleviate the problem of inconsistency between multimodal and unimodal sentiment expression, and provide significant performance advantages.

Key words: multimodal；sentiment analysis：missing modality

多模态情感分析在社交网络、商品评价等应用场景中越来越受到用户的关注，原因是用户越来越倾向于以图文并茂的方式来表达自己的态度和情感。近年来，多模态信息在情感分析中发挥了重要的作用，因为不同模态之间的结构互补性可以提升社交网络中大量多模态信息的情感分析效果[1]。

多模态信息中文本模态通常包含丰富的信息，包括抽象的思想和情感。同时，文本模态还包含了背景和历史等上下文信息，这些信息有助于模型理解用户的意图。视频模态中的色彩和人物表情等信息对情感分析也具有重要作用[2]。因此，如何利用这些信息来提高情感分析的准确性至关重要。

在多模态情感分析中，综合考虑多模态信息对于提高情感分析的准确率至关重要。通常情况下，单一模态的信息并不足以完整地表达用户的情感极性。这是因为用户的情感极性通常是通过不同模态的协同表达来确定的，而每个模态的情感表达可能会有所不同。这也反映了同构信息和异构信息在不同模态之间的存在。因此多模态情感分析需要同时综合考虑这两种信息来做出判断。

然而，在实际应用中，多模态信息并不总是有效存在，尤其是视频模态的缺失情况最为常见。常见的潜在原因包括：当用户的网络环境波动较大时，可能导致视频加载失败；在某些论坛中使用外部链接的视频可能无法访问；摄像头可能被阻挡，或者由于光照过强或过暗而导致视频几乎无法使用。因此，在模态缺失的情况下，有效恢复缺失的模态以进行多模态情感分析是非常必要的。

针对存在模态缺失的多模态情感分析，目前有两种主流方法。一种是显式重构缺失模态，即通过生成新的数据来重建缺失的模态。这种方法可以学习数据的潜在表示，并提高模型在存在模态缺失的情况下的鲁棒性。例如，Vincent 等[3]使用自编码器（auto encoder, AE）来提取输入数据的特征。

另一种方法是通过学习不同模态之间的关系来获得隐式学习模态之间的联合表示。这种方法通过最大程度地保持循环一致性损失，以最小化不同模态之间的信息丢失。例如，IF-MMIN[4]模型通过建立基于中心距离的约束训练策略来预测缺失模态的同构特征。通过引入同构特征，该模型减少了缺失模态异构特征对学习模态之间联合表示的影响。

然而，当前这两种方法主要关注如何在语义空间中引入不同模态之间的同构信息，当存在有模态缺失时，其对应的信息也会丢失，模型无法建立多模态信息间的互补性，不能够充分利用多模态信息间的互补性优势来缓解异构模态之间语义差异带来的多模态与单模态情感表达不一致的问题。因此，在存在缺失模态的条件下，如何对多模态建模是多模态学习的重要问题。

针对上述问题，本文以视频模态缺失作为典型场景，提出了一种显式地恢复缺失模态的同构和异构信息的多模态情感分析模型。该模型采用两阶段缺失模态恢复的方法。首先，通过翻译和联想两个阶段恢复缺失模态的同构和异构特征信息，然后将已知模态和恢复的缺失模态特征进行融合，进行一致性情感分析。在翻译阶段，利用可用模态生成缺失模态的特征。在联想阶段，构建全局视觉词典，并基于全局视觉关联信息逐块恢复视频模态特有的异构信息，从而恢复视觉模态潜在的互补信息，以提高情感分析的一致性。

1 相关工作

1.1 多模态情感分析

多模态任务的核心问题是如何更好地融合不同模态特征以完成建模。目前主要存在以下几种融合方法：

早期融合：在分类开始前融合各个模态特征，这是一种特征层面的融合。早期融合能够有效提取模态之间的交互信息，但可能忽略模态内部的交互信息。早期融合的模型包括 EF-LSTM[5]，该模型通过将不同模态的特征进行拼接得到多模态表示。Nagrani 等[6]引入瓶颈单元，将跨模态流动限制在网络的最后一层，在前面的层中只使用单模态学习。

后期融合：每个模态进行分类后，根据不同模态的分类结果进行加权或投票等方案，作为最终的分类结果。这是一种决策阶段的融合，能够提取模态内部的交互信息，但对模态间的交互信息获取不足。后期融合的模型包括 LF-LSTM[5]，该模型为每个模态特征向量设置了 LSTM 网络，用于单模态编码，最后将它们拼接作为多模态的特征表示。

混合融合：对相关性较弱的模态进行早期融合，而对相关性较强的模态进行后期融合。例如，CentralNet[7]考虑到某些特征在不同模态之间存在的时间关联性，通过在不同模态独立的决策网络中间层联合利用多任务训练进行优化，取得了良好的效果。

以上是当前主要的模态融合方法，它们各有优劣，早期融合能够提取模态间的交互信息，后期融合能够提取模态内的交互信息，而混合融合则结合了两者的优点，但却更加复杂。

1.2 模态缺失的多模态情感分析

Tran 等[8]首次提出并定义了多模态数据中的模态缺失问题。由于不可预见的传感器故障，可能会导致无法获取某些传感器的信息。因此，在分析联合模态数据时，样本的模态覆盖范围可能不同，

这就是所谓的模态缺失。显然，最简单的处理方法是删除存在缺失数据问题的模态或数据项，但这两种方法都会导致有用信息的丢失。另一种方法是根据特征元素间的关联性，从可观察到的元素中推断出缺失元素，即缺失模态的恢复。多模态情感分析中的模态缺失工作可以分为显式重构缺失模态和隐式学习多模态联合表示两种方法。

显式重构缺失模态是通过现有的可用模态恢复缺失模态的特征，进而进行多模态特征融合。例如，变分自编码器(variational auto-encoder, VAE)和生成对抗网络(generative adversarial network, GAN)都可以用作生成符合模态数据实际分布的方法。Cai 等[9]通过 GAN 将模态缺失问题转化为视频生成问题，并在实验中取得了良好的结果。

隐式学习多模态联合表示则尝试根据未缺失模态学习到模态间隐含的联合表示，通过将特征投影到一个共享的子空间中以实现特征对齐和融合[10]。其中一些典型的方法包括多模态循环翻译网络(multimodal cyclic translation network, MCTN)[8]和标签辅助的变形编码器(tag-assisted transformer encoder, TATE)[11]。MCTN 利用循环翻译学习模态之间的联合表示特征，而 TATE 则提出了一种标签辅助的 Transformer 方法，通过使用 Transformer 中的标签编码模块在缺失模态类型和数量不确定的情况下学习多模态潜在表示信息。

现有的方法，无论是显式重构缺失模态还是隐式学习多模态联合表示特征方法，都侧重于使用模态的同构信息，对于模态的异构信息利用不足，无法充分利用异构模态的信息以提升模型性能。因此，本文提出了一种基于两阶段的缺失模态特征恢复方法。

2　任务定义

多模态情感分析任务是给出一组图文模态数据 $\boldsymbol{S}=[\boldsymbol{x}_t,\boldsymbol{x}_v]$ 其中 $\boldsymbol{x}_t$ 和 $\boldsymbol{x}_v$ 分别表示文字模态和视频模态。$\boldsymbol{x}_t=[\boldsymbol{x}_{t1},\boldsymbol{x}_{t2},\ldots,\boldsymbol{x}_{tn}]$，$\boldsymbol{x}_v=[\boldsymbol{x}_{v1},\boldsymbol{x}_{v2},\ldots,\boldsymbol{x}_{vm}]$，其中 n 和 m 分别表示文本和视频模态的长度。对这组图文模态数据 $\boldsymbol{S}$ 对应情感类别进行分类（如积极，中性，消极）。本文研究视频模态 $\boldsymbol{x}_v$ 缺失情况下的多模态情感分析。以三分类任务为例，定义如下：对于存在模态缺失的图文模态数据 $\boldsymbol{S}=[\boldsymbol{x}_t,\boldsymbol{x}_v^{'}]$，其中 $\boldsymbol{x}_v^{'}$ 表示对应模态缺失，对存在模态缺失的图文模态数据 $\boldsymbol{S}$ 进行分类，表示为

$$\mathrm{F}([x_t,x_v^{'}])=e_t,t\in[pos,neu,neg] \tag{1}$$

其中：F 表示多模态情感分类模型，对于给定输入的多模态数据 $[\boldsymbol{x}_t,\boldsymbol{x}_v^{'}]$，通过模型判别得到多模态数据的 e_t，情感类别为积极（pos）、中性（neu）或者消极（neg）。

3　模型描述

本文以视频模态缺失作为模态缺失情况下多模态情感分析典型场景，提出一种基于两阶段缺失模态恢复多模态情感分析模型(a multimodal sentiment analysis model based on two-stage missing modal recovery, TSMR)，该模型旨在解决多模态情感分析中模态缺失的问题，并建模缺失模态的同构信息和异构信息。模型的整体结构如图 1 所示。

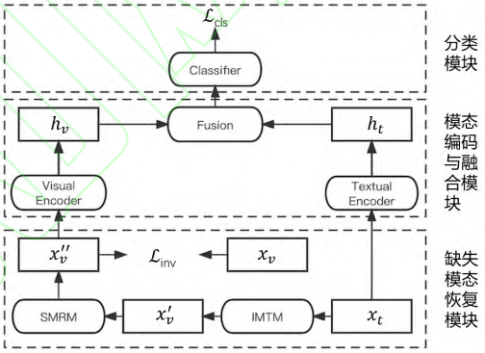

图 1　TAR 模型结构图

Fig.1　TSMR model structure

TSMR 主要分为 3 个模块：缺失模态恢复模块、模态编码与融合模块和分类模块。缺失视频模态恢复模块包括翻译和联想两个阶段。翻译阶段以文本模态原始特征表示 $\boldsymbol{x}_t$ 作为输入，通过不变模态翻译模块(invariant modal translation module, IMTM)子模块生成具有与已知文本模态同构信息的视频模态特征 $\boldsymbol{x}_v^{'}$。联想阶段以翻译阶段输出的视频模态特征 $\boldsymbol{x}_v^{'}$ 作为输入，通过相似模态恢复模块（similar modal recovery module, SMRM）子模块扩展恢复视频模态的异构信息，得到包含同构与异构信息的视频模态特征 $\boldsymbol{x}_v^{''}$。在模型训练阶段，通过计算恢复出的视频模态特征 $\boldsymbol{x}_v^{''}$ 与数据集包含的视频模态特征之间的损失，使恢复生成的视频特征更接近于原始视频真实特征表示。模态编码与融合模块包含模态特征独立编码阶段和特征融合阶段。在文本模态特征编码阶段，利用 LSTM，通过 CNN 网络对视频模态特征进行编码，获得对应的特征向量。在模

态融合阶段，利用 Transformer 中的注意力机制，将缺失视频模态恢复模块输出的恢复生成的视频特征和文本模态特征编码模块输出的特征向量，在模态内和模态间进行融合。情感判别模块使用融合后得到的增强序列 $\boldsymbol{h}^{'}$ 和经过过滤上下文的 $\boldsymbol{h}^{''}$，通过情感分类器进行情感极性判断。

4 TAR 模型

4.1 缺失模态恢复模块

在这一模块中，TSMR 通过 IMTM 子模块将文本模态重建出带有模态间同构信息的视频模态，接着通过 SMRM 子模块从带有模态间同构信息的视频模态中恢复出视频模态的异构信息。最后，将带有模态间同构信息和异构信息的视频模态特征输入到下一模块。

4.1.1 IMTM 模块

鉴于不同模态信息组合联合对应了一致的总体情感极性[11-12]，TSMR 合理假设同一组多模态数据中存在有隐含语义相同的同构信息。因此，为了恢复缺失模态的同构信息，IMTM 模块使用编码器与解码器预测生成缺失模态带有同构信息的特征表示。

将已知模态特征 $\boldsymbol{x}_t$ 送入多头注意力的编码器 E 中得到 $\boldsymbol{x}_e$，接着将被编码的 $\boldsymbol{x}_e$ 再送入到同样是多头注意力的解码器 D 中，获取恢复出的带有同构信息的特征 $\boldsymbol{x}_v^{'}$。

$$x_e = E(x_t) \tag{2}$$

$$\boldsymbol{x}_v^{'} = \boldsymbol{D}(\boldsymbol{x}_e) \tag{3}$$

$$\boldsymbol{K}_{m_i} = \boldsymbol{MHA}_{m_i-1}(\boldsymbol{K}_{m_i-1}, \boldsymbol{K}_{m_i-1}, \boldsymbol{K}_{m_i-1}) \tag{4}$$

其中：E 和 D 都是相同结构的 6 层 4 头注意力，具体计算方法参考文献[18]，每一层是 $\boldsymbol{K}_{m_i}$，当 m 为 t 时，$\boldsymbol{K}_{t_i}$ 是编码器 E 的第 i 层，当 m 为 e 时，$\boldsymbol{K}_{e_i}$ 是解码器 D 的第 i 层。输入 $\boldsymbol{K}_{m_i}$ 是第 i-1 层的输出，当 i 为 1 时，$\boldsymbol{K}_{t_1}$ 是 $\boldsymbol{x}_t$，$\boldsymbol{K}_{e_1}$ 是 $\boldsymbol{x}_e$。

4.1.2 SMRM 模块

考虑到每个模态通常包含与情感表达相关的独特语义信息，这种信息难以用其他模态表示。这些独特的内容是导致多模态情感分析结果不一致的关键因素。TSMR 假设同一组多模态数据中每个模态都具有潜在特异、自有的异构信息。既然这些异构信息是每个模态独有的，为了恢复缺失模态的异构信息，就需要利用缺失模态的可用部分，为恢复模块提供全局的可幻想视频局部参考，TSMR 构建了一个视觉词典（Visual Dictionary，VD）表示为 $\boldsymbol{D} \in \boldsymbol{R}^{k*c}$，其中 k 表示可幻想视频局部向量的数量，c 表示每个向量的维度大小，对于其中的第 j 个向量使用 $\boldsymbol{d}_j$ 表示。在构建视觉词典过程中，TSMR 通过将可幻想视频的局部参考进行聚类，得到多个聚类中心，作为该类别的表示，对于 IMTM 模块获得的视频特征，将其送入到一个 Resnet 网络中得到对应的视觉特征 $\boldsymbol{V}$，对于 $\boldsymbol{V}$ 中的第 i 个特征向量本文使用 $\boldsymbol{v}_i$ 表示，计算视觉特征 $\boldsymbol{v}_i$ 与 $\boldsymbol{d}_j$ 之间的 $L2$ 距离，获取视觉词典 VD 中与视觉特征 $\boldsymbol{v}_i$ 最相似的特征向量表示，

$$\boldsymbol{q}_i = \boldsymbol{d}_{\arg\min_j \|\boldsymbol{v}_i - \boldsymbol{d}_j\|_2} \tag{5}$$

通过利用与视觉特征最相似的特征向量表示 $\boldsymbol{q}_i$，我们可以表示翻译的视频特征 $\boldsymbol{x}_v^{'}$，从而最终更新并获得具有同构信息和异构信息的恢复视频特征 $\boldsymbol{x}_v^{''}$。

$$\boldsymbol{x}_v^{''} = [\boldsymbol{q}_1, \boldsymbol{q}_2 ... \boldsymbol{q}_n] \tag{6}$$

最后使用均方误差（mean-square error，MSE）计算恢复的视频特征与原始视频特征的损失。

$$\boldsymbol{L} = \boldsymbol{MSE}(\boldsymbol{x}_v^{''}, \boldsymbol{x}_v) \tag{7}$$

翻译阶段缺失视频异构信息恢复过程如图 2 所示。视觉目标也就是我们通过 IMTM 模块恢复的具有同构信息视觉表示，对于视觉目标中的每个特征视觉词典中找到与之最相似的特征向量，然后利用在视觉词典对应的相似向量将视频特征表示，从而获得具有视频异构信息的向量表示。

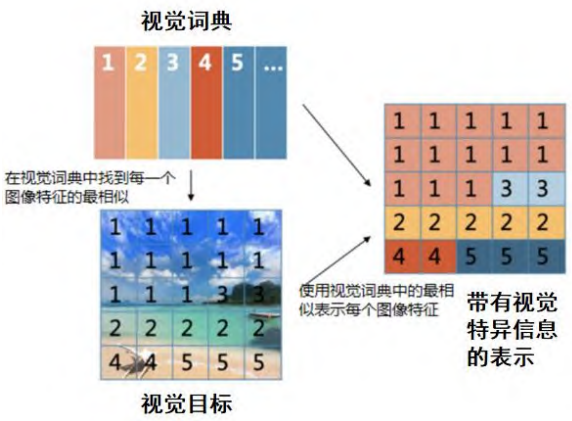

图 2 翻译阶段缺失视频异构信息恢复过程

Fig. 2　Recovery process of heterogeneous information from missing videos in the translation phase

4.2　模态编码与融合模块

完成缺失模态恢复后，将恢复的缺失模态与未缺失模态分别进行编码。

$$\boldsymbol{h}_t = \mathrm{LSTM}(\boldsymbol{x}_t) \tag{8}$$

$$\boldsymbol{h}_v = \mathrm{CNN}(\boldsymbol{x}_v^{''}) \tag{9}$$

为了更好地获取多模态联合表示，TSMR 在模态融合阶段在模态内部和模态之间依次应用注意力机制获取输入序列中每个时间步的动态信息。

接下来通过编码相邻元素信息的 $\boldsymbol{h}_m$ 利用注意力提取模态内信息和模态间信息，如式（6）和式（7），当 n 和 m 不同时表示学习模态间信息，其中查询来自 h_m，键值来自 h_n，当 n 为 t，m 为 v 时分别表示文本与视觉模态，表示提取视觉模态内信息和文本与视觉模态的交互，

$$h_{m\to m} = \mathrm{Transforme\ r}(h_m, h_m, h_m) \tag{10}$$

$$h_{n\to m} = \mathrm{Transforme\ r}(h_m, h_n, h_n) \tag{11}$$

最后，将所有模态内和模态间连接表示 $\boldsymbol{h}$ 通过 Transformer 获得的所有潜在表示连接，作为增强序列 $\boldsymbol{h}_m^{'}$。

$$h = \mathrm{Concat}(h_{m\to m}; h_{n\to m}) \tag{12}$$

$$h_m^{'} = \mathrm{Transforme\ r}(h, h, h) \tag{13}$$

4.3　分类

在分类阶段，TSMR 首先将增强序列输出送入双向 GRU 中，然后将双向 GRU 的输出结果进行一维卷积，接着输入激活函数进行更新，得到更新后的序列 $h_m^{''}$，

$$\boldsymbol{h}_m^{''} = \tanh(\mathrm{BiGRU}(\boldsymbol{h}_m^{'})) \tag{14}$$

$$\boldsymbol{g} = \mathrm{sigmoid}\ (\mathrm{Conv1d}(\boldsymbol{h}_m^{''})) \tag{15}$$

其中：BiGRU 是双向 GRU，Conv1d 是一维卷积，tanh 和 sigmoid 是激活函数。

$$\boldsymbol{h}_m^{'''} = \boldsymbol{h}_m^{''} \otimes \boldsymbol{g} \tag{16}$$

在这里 g 可以被看做是一扇门，用来过滤掉模态融合表示中的不相关上下文，获得过滤不相关内容的表示 $\boldsymbol{h}_m^{'''}$。

最后完成分类，即

$$\hat{\boldsymbol{y}} = \boldsymbol{W}_1 \mathrm{LeakyReLU}\ (\boldsymbol{W}_2 \mathrm{BN}[\boldsymbol{h}_m^{''}, \boldsymbol{h}_m^{'}] + \boldsymbol{b}_2) + \boldsymbol{b}_1 \tag{17}$$

其中：$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 是权重，$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 是偏置。

4.4　算法描述

首先，通过 IMTM 模块通过可用的文本模态获得具有和文本模态同构信息的视觉模态特征。接着，通过 SMRM 模块利用视觉词典从具有文本模态同构信息的视觉模态特征获得具有视觉异构特征信息的视觉模态特征。最后，将具有同构信息与异构信息的视觉模态特征与文本模态特征通过多头注意力机制模态融合，进行分类得到最终的情感类别。

5　实　验

5.1　数据集

本文使用 CMU-MOSI 数据集[14]和 IEMOCAP 数据集[15]评估模型性能。CMU-MOSI 数据集包含有从 YouTube 上收集的 93 个视频，包含 2199 个语句-视频片段，每个片段给出了[-3，3]表示从极端负面到极端正面的 7 种不同情感极性评分。IEMOCAP 数据集包含 10 位演员的 5 个会话，时长总计约 12 h，对应标注有中立、幸福、悲伤、愤怒、惊讶、恐惧、厌恶、挫败、兴奋、其它等 10 种情感。在本文中，实验在 CMU-MOSI 数据集上评估 3 分类([-3, 0), 0, (0,3])模型性能，在 IEMOCAP 数据集上评估 2 分类（积极，消极）模型性能。本文跟从文献[16]利用 OpenFace 提取视频特征表示，跟从文献[17]利用预训练的 Bert 提取文本特征表示。实验参数设置 batch size 为 32，epoch 为 20，隐藏层大小为 300，缺失率 0、0.1、0.2、0.3、0.4、0.5，学习率 0.001，最大文本特征长度 25，最大视觉特征长度 100。

5.2　实验评估

遵循同任务模型评估的方法和原则，采用准确率准确率（A）和 $F1$ 项评价指标对实验结果进行评估，定义为

$$A = \frac{t}{n} \tag{18}$$

其中:t 表示预测正确的数量，n 表示样本总量。

$$\boldsymbol{P} = \frac{\boldsymbol{T}_m}{\boldsymbol{T}_m + \boldsymbol{F}_m} \tag{19}$$

$$\boldsymbol{R} = \frac{\boldsymbol{T}_m}{\boldsymbol{T}_m + \boldsymbol{F}_n} \tag{20}$$

$$F1 = \frac{2PR}{P+R} \tag{21}$$

其中:P 表示精确率，R 表示召回率，T_m 表示预测正确的正例样本，F_m 表示预测错误的反例样本，F_n 表示预测错误的正例样本。

5.3 基线模型

为了验证本方法的有效性，本文使用基线模型MCTN[10]、TransM[13]、TATE[18]、EMMR[11]进行比较。

MCTN：根据机器翻译方法，在模态之间进行翻译来学习多模态鲁棒性联合表示。该方法可以在从源模态翻译到目标模态的过程中捕获到模态之间的联合信息。

TransM：一种端到端的多模态融合方法，它利用注意力机制在模态和编码的多模态特征之间进行翻译。

TATE：基于标签使模型关注到随机模态缺失问题，利用编解码器对缺失模态重构。

EMMR：一种集成多种类编解码器对缺失模态重构的方法。

5.4 实验结果

在 CMU-MOSI 数据集和 IEMOCAP 数据集上的实验结果分别见表 1 和表 2。

表 1 不同模型在 CMU-MOSI 数据集上对比

Tab. 1　Comparison of different models on CMU-MOSI datasets

模型	缺失率											
	0		0.1		0.2		0.3		0.4		0.5	
	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1
MCTN	0.7683	0.5311	0.7611	0.5142	0.7435	0.5007	0.7361	0.4882	0.7317	0.4823	0.6811	0.4576
TransM	0.7966	0.5754	0.7906	0.5721	0.7798	0.5485	0.7131	0.5227	0.7068	0.5114	0.6623	0.4536
TATE	0.7913	0.5698	0.7867	0.5641	0.7629	0.5507	0.7490	0.5483	0.7415	0.5286	0.7292	0.5032
EMMR	0.8073	0.5919	**0.8017**	0.5861	0.7865	0.5688	**0.7801**	0.5521	0.7632	**0.5506**	0.7424	0.5174
TSMR	**0.8141**	**0.5955**	0.7989	**0.5883**	**0.7940**	**0.5807**	0.7789	**0.5613**	**0.7736**	0.5454	**0.7538**	**0.5361**

表 2 不同模型在 IEMOCAP 数据集上对比

Tab. 2　Comparison of different models on IEMOCAP datasets

模型	缺失率											
	0		0.1		0.2		0.3		0.4		0.5	
	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1	准确率	F_1
MCTN	0.7779	0.7024	0.7612	0.6916	0.7357	0.6657	0.7321	0.6634	0.7223	0.6488	0.7151	0.6337
TransM	0.7736	0.7051	0.7649	0.6875	0.7562	0.6739	0.7437	0.6712	0.7297	0.6494	0.7327	0.6310
TATE	0.8082	0.7261	0.8053	0.7156	0.7927	0.7034	0.7827	0.7076	0.7781	0.6871	0.7751	0.6708
EMMR	**0.8174**	0.7497	0.8062	0.7385	0.7989	0.7288	0.7893	0.7122	0.7851	0.7118	0.7804	0.6933
TSMR	0.8123	**0.7644**	**0.8073**	**0.7556**	**0.8017**	**0.7472**	**0.7974**	**0.7435**	**0.7915**	**0.7237**	**0.7838**	**0.7122**

本文在测试数据集中将图片模态的缺失率设置为 0~0.5 不同等级，并以实验结果中第一行的数字表示。其中，0.5 表示多模态数据中视频模态数据缺失率为 50%。从实验结果可以看出，当缺失率提高时，所有模型的性能都会下降。这是因为模态的缺失增加导致多模态数据与情感之间的关联信息量减少、数据分布偏差增大，从而使模型难以学习到鲁棒的分类边界。

在 MOSI 数据集和 IEMOCAP 数据集上，可以观察到除了少数情况下，本文提出的模型均表现出最佳的性能。MOSI 数据集在缺失率为 0.1 的条件下 TSMR 模型没有取得最佳性能是因为在模态较为完整的条件下，本文模型对于模态异构信息的学习效果不足，而 EMMR 预训练网络则能够对整个模型做

出更好的引导。相对于 MOSI 数据集，在 IEMOCAP 数据集上 TSMR 模型的效果更为稳定，仅在 0 缺失率条件下不是最优。从数据集角度分析是 IEMOCAP 数据集是由有限的演讲者在特定话题下演绎不同情感，而 MOSI 数据集具有更多的演讲者和更多的背景和主题，这代表 MOSI 数据集的模态异构信息空间更大，模型要学习 MOSI 数据集中的模态异构信息更困难，因此 TSMR 模型在 MOSI 数据集中缺失率为 0.3、0.4 的时候效果逊于 EMMR 模型，相比于 IEMOCAP 数据集模型在 MOSI 数据集上的效果更不稳定。但总的来说，在大多数情况下，本文模型展现出最佳的实验表现，这说明了本文模型的优越性。

在缺失率较高的条件下，MCTN 和 TransM 的性能下降较快。这说明在模态缺失比例较高的情况下，基于翻译的方法仅能建模跨模态同构信息，而难以学习到缺失模态的异构信息。这证明了本文模型对于模态异构信息的学习能力，能够有效缓解整体多模态与单模态情感表达不一致的问题。

5.5 消融实验

为了更好地探究在本模型中模块对于学习模态异构信息的作用，设计了关于 SMRM 模块的消融实验，在移除了 SMRM 模块后本文将 IMTM 模块的输出直接送入到视觉编码器中并与文本模态进行模态融合。消融实验结果见表 3。根据表 3 可以看到移除 SMRM 模块后模型的性能下降是较为明显的，这说明了 SMRM 模块对于模型学习模态异构信息的有效性。此外，可以看到在 MOSI 数据集中移除 SMRM 模块后，模型的性能变化明显小于 IEMOCAP 数据集，考虑到 IEMOCAP 数据集具有指定的话题和有限的演讲者，而 MOSI 数据集则是从 YouTube 收集的更多演讲者和不特定的话题内容，这意味着 IEMOCAP 的视觉模态异构信息具有更高的相似性，SMRM 模块可以更好的学习到这种更相似的异构信息，而这在 MOSI 数据集中更加困难，因此 SMRM 模块的影响在 IEMOCAP 数据集中更大。

表 3　消融实验结果

Tab. 3　Experimental results of ablation study

模型	CMU-MOSI		IEMPCAP	
	准确率	F_1	准确率	F_1
w/o SMRM	0.8023	0.5767	0.7838	0.7307
TSMR	**0.8141**	**0.5955**	**0.8123**	**0.7644**

6　结　语

本文提出了一种基于两阶段缺失模态恢复多模态情感分析模型 TSMR。该模型通过 IMTM 模块学习模态的同构信息，通过 SMRM 模块学习模态的异构信息，并利用 Transformer 完成模态间的融合。通过注意力机制对模态与模态之间进行计算，充分考虑了模态间的相互关系。

通过在两个数据集上的实验验证，证明了 TSMR 模型同时建模缺失模态的同构信息和异构信息的必要性，并展示了本文模型架构以及翻译-联想两阶段缺失模态信息恢复方法的有效性。特别是在缺失率较高的情况下，本文模型的实验结果强调了缺失模态的异构信息在多模态情感分析中的重要性。在电商平台或者社交网络中对于用户的回复中可能常常存在有模态缺失的情况，这种条件下对于用户对商品的喜恶和对舆论的态度判别有潜在的应用价值。

参考文献：

[1] LI S, DU C, HUANG Y, et al. Modality complementariness: towards understanding multi-modal robustness[C]//ICLR. Proceedings of International Conference on Learning Representations. Kigali: ICLR, 2023: 1-15.

[2] HENDRICKS L A, NEMATZADEH A. Probing image-language transformers for verb understanding[EB/OL]. [2023-10-11]. https://doi.org/10.48550/arXiv.2106.09141.

[3] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Conference on Machine Learning. New York: ICML, 2008: 1096-1103.

[4] ZUO H, LIU R, ZHAO J, et al. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities[C]//ICASSP. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes: ICASSP, 2023: 1-5.

[5] ZADEH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension[C]//AAAI. Proceedings of the 32th AAAI Conference on Artificial Intelligence. California：AAAI, 2018: 5642–5649.

[6] NAGRANI A, YANG S, ARNAB A, et al. Attention bottlenecks for multimodal fusion[J]. Advances in neural

information processing systems, 2021, 34: 14200-14213.

[7] VIELZEUF V, LECHERVY A, PATEUX S, et al. CentralNet: a multilayer approach for multimodal fusion[C]//ECCV. Proceedings of the European Conference on Computer Vision. Munich: ECCV Workshops, 2018: 07275.

[8] TRAN L, LIU X, ZHOU J, et al. Missing modalities imputation via cascaded residual autoencoder[C]//IEEE. Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 1405-1414.

[9] CAI L, WANG Z, GAO H, et al. Deep adversarial learning for multi-modality missing data completion[C]//ACM. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. New York：ACM, 2018: 1158-1166.

[10] NORMAND R, DU W, BRILLER M, et al. Found in translation: a machine learning model for mouse-to-human inference[J]. Nature methods, 2018, 15(12): 1067-1073.

[11] ZENG J, ZHOU J, LIU T. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities[C]//EMNLP. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: EMNLP, 2022: 2924-2934.

[12] HAZARIKA D, ZIMMERMANN R, PORIA S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis[C]//ACM. Proceedings of the 28th ACM international conference on multimedia. New York：ACM, 2020: 1122-1131.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2023-10-11]. http://www.aiotlab.org/teaching/intro2ai/slides/10_attention_n_bert.pdf.

[14] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages[J]. IEEE intelligent systems, 2016, 31(6): 82-88.

[15] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42: 335-359.

[16] 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]//ACM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. New York： ACM, 2018: 2236-2246.

[17] DEVLIN, JACOB et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// NAACL-HLT. Proceedings of NAACL-HLT. Minneapolis: NAACL-HLT, 2019: 4171-4186.

[18] ZENG J, LIU T, ZHOU J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities[C]//ACM. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM, 2022: 1545-1554.