

基于双向分层注意力的多模态情感补偿网络

邱锦宏¹ 李泽南¹

QIU Jinhong LI Zenan

摘要

针对多模态情感分析中模态缺失以及主流注意力机制没有考虑模态间的固有相关性,导致的特征偏差问题,基于此,文章提出了一个基于双向分层注意力的补偿网络。首先对3个模态构建模态关系先验矩阵确定模态间的互补关系,然后设计基于模态关系先验矩阵的双向分层注意力机制,实现缺失模态的特征补偿。实验表明,在 MOSI 和 CH-SIMS 数据集上,当模态缺失率为 50% 时,情感识别准确率分别达到 79.6% 和 78.72%,较最优基线提高 4.34% 和 4.61% 个百分点。

关键词

多模态情感分析; 模态缺失补偿; 分层注意力; 双向注意力机制

doi: 10.3969/j.issn.1672-9528.2025.10.039

0 引言

近年来,随着社交多媒体平台的快速发展,包含多种模态的视频已成为主要的信息载体,这些平台上发布的视频博客或口头评论,包含了用户情感的表达。例如,描述用户谈论产品或电影的视频,视频不仅提供文本信息,还提供丰富的视觉和音频信息。多模态情感分析通过利用视觉和音频中可用的额外信息来预测人类情感,以协助基于文本的预测。

近年来,中外学者在多模态情感分析与识别上取得了一系列突破。张添植等人^[1]提出一种基于文本和图像门控融合机制的多模态方面级情感分析模型,该模型在文本和图像进行交互的同时引入了从数据集图像中提取的形容词-名词对,并将其中形容词的加权作为图像辅助信息,在特征融合阶段,通过构建一种动态控制图像和图像辅助信息输入的的门控机制实现多模态特征融合。余本功等人^[2]提出一种多模态协同对比学习的方面级情感分析模型。该模型在文本侧利用方面词与句子中方面词编码的相似性,在图像侧利用图像经过随机裁剪后在不同顺序下编码的相似性,分别构造对比学习所需的正负样本,进而利用对比学习任务的损失函数,学习到更具区分度的特征表示。Zadeh 等人^[3]提出的张量融合网络(TFN)首先通过3个模态嵌入子网络分别处理文本、视觉和音频模态的输入,然后使用张量融合层,外积运算显式地建模单模态、双模态和三模态之间的交互从而捕捉不同模态间的高阶相关性来提高模型的表现力。孙斌等人^[4]提出一种基于跨模态联合编码的多模态情感分析模型。该模型利用跨模态注意力模块获得两模态的联合特征,并使用联合特征分别调整单模态特征的权重,然后将两者拼接获得多模态融合

特征。

尽管已有研究取得显著进展,但在模态缺失场景下现有方法仍然面临严峻挑战,现有这些方法多采用均匀注意力分配策略,未能充分考虑模态间非对称互补特性。现有研究存在两个关键局限:

一是传统方法如 TFN、LMF^[5]等假设各模态完整可用,当出现传感器故障导致模态缺失时,识别性能骤降;

二是主流的注意力机制没有考虑模态间的固有相关性,导致特征补偿偏差。

为此,本文提出基于双向分层注意力的多模态情感补偿网络:(1)构建模态关系先验矩阵,确定语音、文本和视觉之间的互补强度;(2)设计双向分层注意力机制,通过双向分层注意力来实现缺失特征的定向重构。

1 实验数据集与研究方法

1.1 数据集构成

本文采用的数据集是网络公开数据集 MOSI^[6]和 CH-SIMS^[7],MOSI是由卡内基梅隆大学发布的多模态情感分析数据集。该数据集包含2199个意见视频片段,每个视频都标注了情感强度,情感强度范围为[-3, 3],其中-3表示非常消极,3表示非常积极;CH-SIMS是一个细粒度标注的中文多模态情感分析数据集,包含2281个视频片段,主要特点是同时具有单模态和多模态的情感标签。

1.2 整体架构设计

本文提出的双向分层注意力多模态情感补偿网络包含3个核心模块:

多模态编码模块:文本为使用 BERT 的 [CLS] 标签进行编码,语音为使用独立的 Transformer 编码器进行编码,视觉

为使用独立的 Transformer 编码器进行编码。

双向分层补偿模块：基于模态关系矩阵的双向分层注意力补偿。

情感预测模块：使用多层感知机（MLP）进行情感预测。

架构图：左侧为多模态编码模块；中部为双向分层补偿模块；右侧为情感预测模块。如图 1 所示。

1.3 模态关系矩阵构建

定义模态集合 $M = \{T, A, V\}$ ，通过统计模态间 Pearson 相关系数构建先验矩阵 $R \in \mathbf{R}^{3 \times 3}$ 表达式为：

$$R_{ij} = \frac{1}{N} \sum_{n=1}^N \frac{\text{cov}(X_i^{(n)}, X_j^{(n)})}{\sigma_{X_i} \sigma_{X_j}} \quad (1)$$

1.4 双向分层注意力机制

本文在传统的注意力计算基础上，引入模态关系矩阵作为先验知识。得到改进后的注意力表达式为：

$$\alpha_{ij}^{(h)} = \text{softmax} \left(\frac{Q_i^{(h)} (K_j^{(h)})^T}{\sqrt{d_h}} + \gamma R_{ij} \right) \quad (2)$$

式中： $h \in 1, 2, \dots, H$ 为注意力头索引； Q_i 和 K_j 分别为查询向量和键向量； d 为特征维度； γ 为可学习缩放因子； R_{ij} 为模态关系先验权重。

通过这种计算方式，注意力机制不仅考虑了当前模态间的直接关联，还融入了模态间的先验补偿关系，平衡了注意力得分中原生注意力和模态关系先验的贡献，使得在模态缺失时能够更准确地从其他模态中获取有用信息。之后将多头注意力进行聚合公式从而得到加权的特征表示，表达式分别为：

$$\text{Attn}_i = \text{Concat}(\text{softmax}(\alpha_{ij}^{(1)}), \text{softmax}(\alpha_{ij}^{(2)}), \dots, \text{softmax}(\alpha_{ij}^{(H)})) \quad (3)$$

$$C_i = \text{Attn}_i \cdot [V_j^{(1)} || V_j^{(2)} || \dots || V_j^{(H)}] \quad (4)$$

式中： Attn_i 是聚合后的多头注意力； V_j 是值向量； C_i 是使用注意力加权后的特征表示。

最后对加权特征表示进行残差连接与层归一化得到补偿后的特征，表达式为：

$$X'_i = \text{LayerNorm}(X_i + C_i) \quad (5)$$

在对缺失模态进行补偿的过程中不只需要考虑通过完成模态得到缺失模态，还应该同时考虑对得到的补偿特征进行噪声抑制，需要对特征进行修正，如在利用视觉和音频特征对文本特征进行补偿的时候，抑制与补偿文本无关的面部动作（如眨眼）修正视觉特征，降低与补偿文本冲突的语调成分修正音频特征，即对每对模态 (i, j) 进行双向补偿。

正向补偿主要作用为恢复缺失模态，表达式为：

$$X_j^{\text{comp}} = \text{CrossAttn}(Q_j, K_i, V_i) \quad (6)$$

逆向补偿主要作用为防止噪声传播，表达式为：

$$X_i^{\text{comp}} = \text{CrossAttn}(Q_i, K_j, V_j) \quad (7)$$

最后对两个补偿特征进行融合得到最终的双向补偿融合特征，表达式为：

$$X_{\text{final}} = \beta X_j^{\text{comp}} + (1 - \beta) X_i^{\text{comp}} \quad (8)$$

式中： β 为可学习参数。

1.5 多模态融合与情感预测

在完成模态补偿之后，模型需要将处理后的多个模态特征进行融合，以实现最终的情感预测。

特征融合：将双向分层跨补偿模块得到的补偿特征，以及未缺失特征分解为具有不同投影仪的相似特征和不相似特征。每个投影仪由归一化层、tanh 激活的线性层和 dropout 层组成，然后对其进行拼接，形成融合特征向量。该向量包

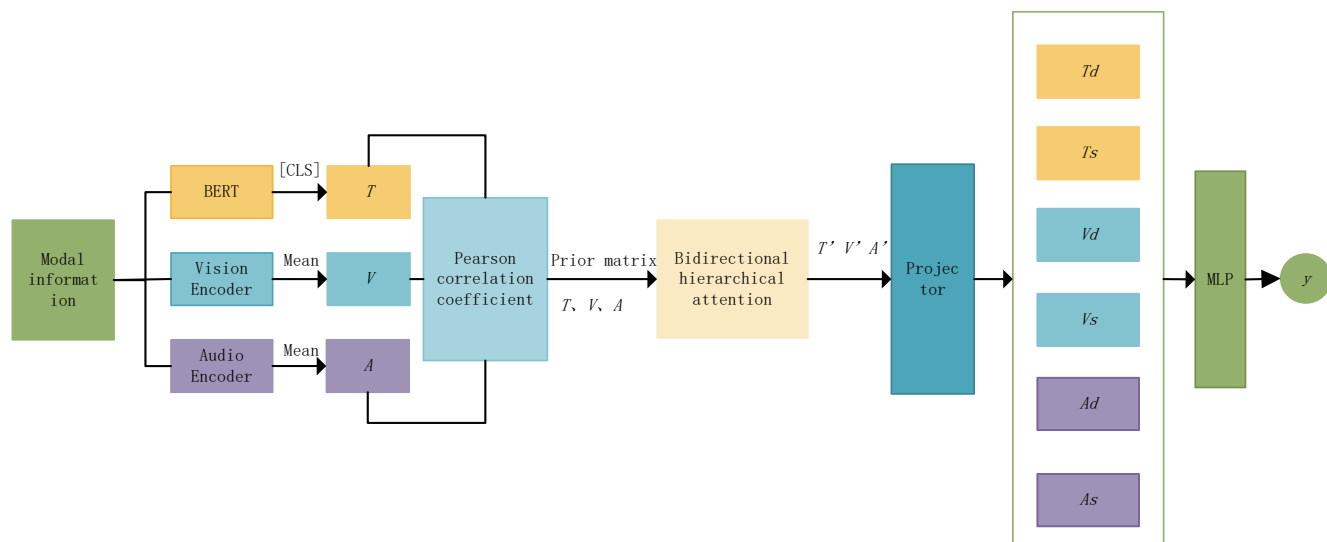


图 1 整体架构图

含了来自不同模态的丰富信息，既包括直接观测到的特征，也包括通过补偿和得到的潜在特征。

情感预测：将融合特征向量输入到情感预测模块，采用多层感知机（MLP）进行回归。情感预测模块根据融合特征向量中的信息，学习到情感与特征之间的映射关系，从而能够对新的样本进行情感预测。

模型训练与优化：在模型训练过程中，通过反向传播算法优化总损失函数，不断调整模型的参数，使得模型在训练数据上能够准确地预测情感标签。同时，通过验证集对模型进行评估和调整，避免过拟合，提高模型的泛化能力。

2 实验结果与分析

2.1 实验环境与配置

在本文研究中实验环境：Ubuntu 20.04.6 LTS 操作系统、Intel(R) Xeon(R) Gold 6248R CPU@3.00 GHz CPU、251.5 GB 内存；开发语言为 Python3.8.8；开发软件为 Visual Studio Code；CUDA 版本为 11.1；对于 MOSI 数据集，实验的 batch-size 为 16，实验设置的 epoch 为 25，学习率为 1e-4；对于 CH-SIMS 数据集，实验的 batch-size 为 32，实验设置的 epoch 为 50，学习率为 1e-4。

2.2 结果与分析

2.2.1 消融实验

基于上述的实现环境和配置下，本文对改进前后的模型在两个数据集上分别进行了消融实验，改进后的模型和改进前的模型的实验结果对比如表 1 所示，以原模型 ConFEDE^[8]为基线，依次添加正向补偿 Pcomp 以及逆向补偿 Rcomp 的模型，分别在 MOSI 和 CH-SIMS 两个数据集缺失比例为 0.1、0.3 和 0.5 情况下的实验结果。

表 1 改进前后实验结果对比

数据集	Model	单位：%					
		0.1		0.3		0.5	
		Acc-2	F_1	Acc-2	F_1	Acc-2	F_1
MOSI	ConFEDE	81.49	81.52	79.59	79.64	78.13	78.19
	ConFEDE + Pcomp	82.65	82.69	80.47	80.33	79.12	79.23
	ConFEDE + Pcomp + Rcomp	83.38	83.46	81.55	81.45	79.60	79.59
CH-SIMS	ConFEDE	80.74	80.54	78.56	78.82	77.46	76.51
	ConFEDE + Pcomp	81.40	81.30	79.21	79.35	78.43	78.36
	ConFEDE + Pcomp + Rcomp	81.84	81.62	79.87	79.46	78.72	78.65

对于 MOSI 数据集，在 0.1 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 1.16% 和 1.17%，完整模型 Acc-2 和 F_1 分别提高了 1.89% 和 1.94%；在 0.3 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 0.88% 和 0.69%，完整模型 Acc-2 和 F_1 分别提高了 1.96% 和 1.81%；在 0.5 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 0.99% 和 1.04%，完整模型 Acc-2 和 F_1 分别提高了 1.47% 和 1.4%。

对于 CH-SIMS 数据集，在 0.1 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 0.66% 和 0.76%，完整模型 Acc-2 和 F_1 分别提高了 1.1% 和 1.08%；在 0.3 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 0.65% 和 0.53%，完整模型 Acc-2 和 F_1 分别提高了 1.31% 和 0.62%；在 0.5 缺失比例的情况下，只添加正向补偿的模型 Acc-2 和 F_1 分别提高了 0.97% 和 1.85%，完整模型 Acc-2 和 F_1 分别提高了 1.26% 和 2.14%。

仅添加正向补偿对模型性能有一定的提升效果，而完整模型在所有缺失比例下均表现出更优的性能，特别是在高缺失比例（0.5）时，完整模型在 CH-SIMS 数据集上的 F_1 指标提升更为显著，这表明正向和逆向补偿的结合不仅补偿了缺失模态，还有效去除了噪声，显著提升了模型的鲁棒性和性能。

2.2.2 对比实验

本文用 TransM^[9]、MMIN^[10] 和 TATE^[11] 模型作为对比模型，在表 2 中可以看到本文模型在两个数据集的不同比例的缺失情况下都优于 3 个模型，在 0.1 缺失比例情况下，本文模型在 MOSI 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 0.48% 和 1.02%，在 CH-SIMS 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 0.29% 和 0.27%，

在 0.3 缺失比例情况下，本文模型在 MOSI 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 1.01% 和 1.26%。在 CH-SIMS 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 1.16% 和 0.83%，在 0.5 缺失比例情况下，本文模型在 MOSI 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 4.34% 和 4.44%，在 CH-SIMS 上，对比最优基线模型 TATE，Acc-2 和 F_1 分别提高了 4.61% 和 4.62%。

表 2 不同网络模型实验结果对比

数据集	Model	单位: %					
		0.1		0.3		0.5	
		Acc-2	F_1	Acc-2	F_1	Acc-2	F_1
MOSI	TransM	79.63	79.34	72.89	72.55	68.04	67.43
	MMIN	81.52	81.63	79.14	79.08	70.46	70.37
	TATE	82.90	82.44	80.54	80.19	75.26	75.15
	本文模型	83.38	83.46	81.55	81.45	79.60	79.59
	TransM	78.10	78.02	71.66	71.58	67.20	67.19
CH-SIMS	MMIN	80.46	80.29	77.82	77.56	69.75	69.47
	TATE	81.55	81.35	78.71	78.63	74.11	74.03
	本文模型	81.84	81.62	79.87	79.46	78.72	78.65

3 结论与展望

为了改善多模态情感分析面临的模态缺失现象导致的识别性能下降，以及现有的注意力机制没有考虑模态间的固有相关性，导致特征补偿偏差，本文提出基于双向分层注意力的多模态情感补偿网络，通过双向分层注意力有效解决了模态缺失场景下的情感分析难题。利用双向分层注意力对多模态情感分析的缺失问题做了一定的探索，但因为需要利用现有模态来补偿缺失模态，从而增加了设备计算力的需求，之后主要研究方向为如何减小模型所需算力。

参考文献:

[1] 张添植,周刚,刘洪波,等. 基于文本和图像门控融合机制的多模态方面级情感分析[J]. 计算机科学,2024,51(9):242-249.

[2] 余本功,邢钰,张书文. 多模态协同对比学习的方面级情感分析模型[J]. 数据分析与知识发现,2024,8(11):22-32.

[3]ZADEH A, CHEN M H, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing(EMNLP).Brussels:ACL,2017: 1103-1114.

[4] 孙斌,江涛,贾莉,等. 基于跨模态联合编码的多模态情感分析[J]. 计算机工程与应用,2024,60(18):208-216.

[5] 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.Brussels: ACL, 2018: 2247-2256.

[6] ZADEH A, ZELLERS R, PINCUS E, et al. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in

online opinion videos[J]. IEEE intelligent systems, 2016, 31(6): 82-88.

[7] YU W M, XU H, MENG F Y, et al. CH-SIMS: a chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proceedings of the 58th annual meeting of the association for computational linguistics,Brussels:ACL,2020:3718-3727.

[8] YANG J D, YU Y K, NIU

D, et al. ConFEDE: contrastive feature decomposition for multimodal sentiment analysis[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Brussels:ACL,2023: 7617-7630.

[9] WANG Z L, WAN Z H, WAN X J. Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis[C]//WWW '20: Proceedings of The Web Conference 2020.NewYork:ACM,2020: 2514-2520.

[10] ZHAO J M, LI R C, JIN Q. Missing modality imagination network for emotion recognition with uncertain missing modalities [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Brussels:ACL,2021:2608-2618.

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

【作者简介】

邱锦宏（2002—），男，江西上饶人，硕士研究生，研究方向：机器学习与自然语言处理。

（收稿日期：2025-05-12 修回日期：2025-09-29）