LAOIS：多层次融合的图文贴纸情感分析模型

李 伟[1] 王东娟[1]
LI Wei WANG Dongjuan

摘 要 随着社交媒体和聊天应用的普及，多模态的图文数据已成为用户情感表达的主要途径之一。由于图文数据在空间上呈现独立性，但又具有明显的语义关联，现有方法忽略了模态之间的双向关联性和互补性，导致模型提取的特征无法全面反映多模态数据的内在关联。针对上述问题，文章提出了多层次特征融合的贴纸情感分析模型，在浅层通过对比学习策略，实现模态之间的信息共享和转移，深度挖掘样本之间的语义相关性；在深层将不同模态的信息结合，提高模型对多模态数据的整体抽象能力，捕捉模态间的互补性；多层次的特征融合提取出具有双向关联性和互补性的综合特征用于情感分析。在公开数据集 MVAS-single 上的情感分类准确率和 F_1 值分别提高 2.44% 和 1.05%，验证了该方法的有效性。

关键词 图文情感分析；特征提取；多模态特征融合；对比学习

doi：10.3969/j.issn.1672-9528.2024.12.042

0 引言

近年来，随着社交媒体和聊天应用的普及，人们在日常交流中用来表达情感的方式不再局限于单一形式，由于纯文本的交互难以反映用户的真实情感，多模态的图文、音视频等数据以其生动、直观的特点，逐渐成为一种流行的情感表达方式[1]，其中最常见的形式为图像嵌入文本信息相结合[2]。

近年来，情感分析领域的研究收获颇丰，尤其是在图像与文本关联性、模态间互补性以及情感分布不均匀性方面，均取得了众多新成果。然而，现有研究往往集中在某一特定方面，而较少综合考虑这些问题之间的关系。Zhou 等人[3]通过联合训练图像和文本的表示，实现更好的图像与文本匹配性能；Peng 等人[4]通过生成网络和判别网络（GAN）的协同训练，使得模态之间的信息更好地互相补充，实现模态间的互补性学习；Li 等人[5]通过引入基于注意力的视觉处理网络，解决了情感分布的不均匀性。当前研究虽然取得了一些进展，但在综合考虑图像与文本关联性、模态间互补性和情感分布不均匀性方面仍存在局限。本文提出一种基于多层次融合的图文贴纸情感分析模型，通过多层次融合文本与图像信息，充分挖掘贴纸数据的双向关联性和互补性，从而提升情感判别能力。

1 相关研究

在表情贴纸情感识别的研究领域，与其密切相关的研究主要有文本的情感分析、图像的情感分析和图文结合的情感分析。随着社交媒体的快速发展，越来越多的用户选择通过文本、图片或视频等多种方式来表达自己的情感和看法，研究者们开始将图像和文本相结合进行分析。You 等人[6]提出联合图像和文本的情感识别方法，在 imagenet 上使用预训练的 CNN 模型进行微调，进而用来提取视觉特征；对于文本特征，使用一种无监督语言模型来学习文档和段落的分布式表示；Chen 等人[7]提出了一个名为 GME-LSTM（A）的模型，旨在融合图像和文本信息，该模型基于门控多模态嵌入 LSTM，缓解了由于模态之间存在噪声导致的融合困难问题；Dai 等人[8]在 2020 年引入情感嵌入学习的概念，旨在将情感信息映射到一个共享的嵌入空间，使得在这个空间中的情感表示能够在不同的模态之间进行迁移和共享；2022 年，Hu 等人[9]在模态和样本之间引入对比学习，更好地捕捉情感和情感之间的差异和一致性。当前的研究聚焦于图像与文本关联性、模态间互补性和情感分布不均匀性等关键方向，力图深化对多模态情感分析问题的理解，并为实际应用中的情感分析提供更强大的解决方案。

2 多层次融合模型（LAOIS）

前述研究针对模态间融合的困难提出了多种解决方法，但通常集中在特定问题上，忽视了模态之间的双向关联性和互补性。基于这些研究，本文提出了一种多层次融合的图文贴纸情感分析模型（LAOIS），旨在提高情感分类的准确性和鲁棒性。该模型主要由图文特征提取模块、浅层特征融合模块和深层特征融合模块组成，形成了多层次融合模块（ML-F），如图 1 所示。

1. 三峡大学计算机与信息学院 湖北宜昌 443002

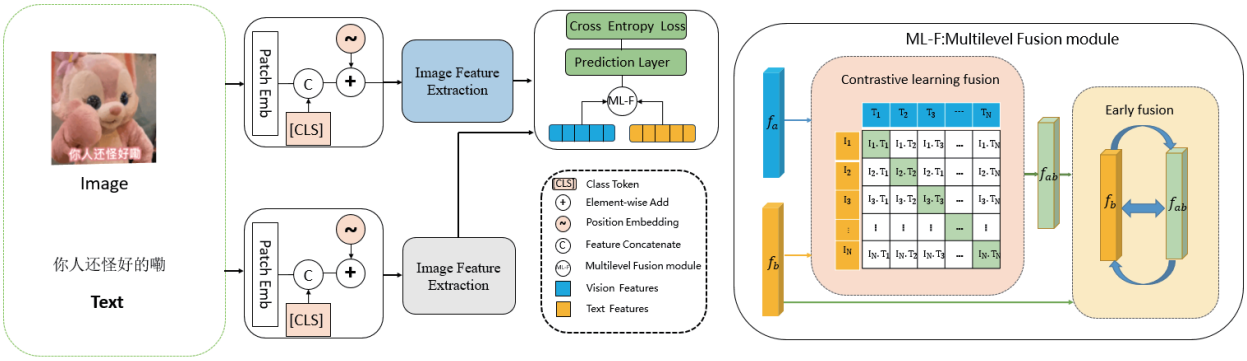

图 1 基于对比学习及早期特征交互的多层次融合模型框架

2.1 浅层时序特征融合

浅层特征融合模块的输入为单文本特征向量、单图像特征向量。对于文本特征的提取，首先对文本进行预处理，将文本进行分词并去除无意义字符，将预处理的文本传递给 BERT 模型，由 bert 的 Encoder 层输出文本特征向量 $\boldsymbol{F}^T \in \mathbf{R}^{1\times C}$。

对于图像特征的提取，使用了预训练的 BLIP[10] 模型，数据通过 BLIP 的 Image-encoder 模块输出视觉特征向量 $\boldsymbol{F}^I \in \mathbf{R}^{1\times C}$，$C$ 是编码器的嵌入维数，用于从表情贴纸图像中提取视觉特征。

将提取到的文本和图像特征映射到同一个共享的情感表示空间中。通过最大化正样本（图像和相应文本）之间的相似性，同时最小化负样本（图像和不相应的其他文本）之间的相似性。从而实现模态之间的信息共享和转移。

2.2 深层时序特征融合

在深层时序的融合中，计算复杂度相对较低的早期融合方法有利于增强模型对多模态数据的整体理解和表达能力，通过在信息处理的早期阶段整合不同模态的数据，能够更全面地捕捉多模态数据之间的互补性。文本特征 $\boldsymbol{F}^T \in \mathbf{R}^{1\times C}$ 的提取基于 BERT 模型，这使模型能够深入挖掘文本中的语义信息。同时，对比学习提取的综合特征 $\boldsymbol{F}^C \in \mathbf{R}^{1\times C}$ 则突出了图像和文本之间的双向联系，强调了它们在情感表达中的关联性。将文本特征 $\boldsymbol{F}^T$ 和综合特征 $\boldsymbol{F}^C$ 通过 Concat 操作连接起来，得到包含关联性与互补性的综合特征表示 $\boldsymbol{f}_c$ 作为全连接层的输入，公式为：

$$\boldsymbol{f}_c = \text{concat}\left(\boldsymbol{F}^T, \boldsymbol{F}^C\right) \tag{1}$$

对融合后的特征 $\boldsymbol{f}_c$ 用于最终的情感识别，公式为：

$$\emptyset = \text{fusion}\left(\boldsymbol{f}_c, \boldsymbol{\theta}_f\right), \emptyset \in \mathbf{R} \tag{2}$$

式中：ϕ 代表模型预测的情感标签，$\boldsymbol{\theta}_f$ 代表全连接层中的参数。

2.3 ML-F 多层次特征融合模块

ML-F 模块设计基于两个关键子模块：浅层时序特征融合和深层时序特征融合。浅层子模块采用对比学习策略，深层子模块则采用早期融合方法，以实现图像与文本特征的有机融合。

在浅层时序特征融合模块中，输入特征向量包括文本特征向量 $\boldsymbol{F}^T$ 和视觉特征向量 $\boldsymbol{F}^I$，这两者由不同编码器独立提取，以保持在空间上的相互隔离，避免数据交叉引起噪声。浅层时序特征融合模块的公式为：

$$\boldsymbol{F}^T = \text{BERT(text)} \qquad \boldsymbol{F}^I = \text{BLIP(image)} \tag{3}$$

然后，浅层时序特征融合模块将输入的 $\boldsymbol{F}^T$ 与 $\boldsymbol{F}^I$ 特征映射到同一个共享的空间中：

$$\boldsymbol{h}^1 = \text{Linear}(\boldsymbol{F}^T) \qquad \boldsymbol{h}^2 = \text{Linear}(\boldsymbol{F}^I) \tag{4}$$

通过对比损失函数使得相似的图像和文本在嵌入空间中的表示更加接近，而不相似的图像和文本在该空间中的表示距离更远，模型能够初步学习到具有语义关联性的综合特征 $\boldsymbol{F}^C$：

$$\boldsymbol{F}^C = \text{ContrastiveLoss}\left(\boldsymbol{h}^1, \boldsymbol{h}^2\right) \tag{5}$$

深层时序的特征融合将对比学习输出的特征作为深层时序融合模块的输入，并与文本特征向量 $\boldsymbol{F}^T$ 再次融合：

$$\boldsymbol{f}_c = \text{concat}\left(\boldsymbol{F}^T, \boldsymbol{F}^C\right) \tag{6}$$

浅层特征融合模块中对比学习的策略解决了模态之间存在噪声导致的融合困难问题；深层特征融合模块中早期融合的方法，挖掘了多模态数据之间的互补性，捕捉与情感紧密相关的联合特性。

3 实验与结果分析

3.1 数据集

为评估本文提出的方法在表情贴纸情感识别任务上的性能，进行了一系列对比实验，如表 1 所示。

表 1 实验数据集统计

名称	总数	中性	积极	消极
Our-dataset	3008	315	1019	1674
MVSA-single	4511	470	2683	1358

由于公开且带有真实情感标签的表情贴纸样本稀少，本文利用相关文献 [11] 中提出的思路建立了小型图文结合的表情贴纸分析数据集。另外对 MVSA-Single 数据集进行了预处理，在 MVSA-Single 数据集上对该模型进一步验证。

3.2 实验设置

实验分别采用 bert 和 clip 对文本与图像进行特征提取。文本数据使用预训练的 BERT 模型进行参数初始化，词向量维度设为 768。由于贴纸数据的文本描述多为短文本，句子的最大长度设置为 20，以提高运算速率。图像模态的输入像素均为 224×224。学习率设置为 5e-3，batch 大小为 16，优化函数选用 Adam。为防止过拟合，使用 Dropout 正则化层，通过随机丢弃部分神经网络单元来降低模型复杂度，提高泛化能力。

3.3 对比方法

表 2 展示了不同多模态图文融合情感分类方法的实验结果比较。本文提出的模型相较于 MultiSentiNet[12] 方法在 MVSA-single 数据集上的准确率和 F_1 值分别提升了 3.22% 和 4.73%，由于 EF-CapTrBERT 和 MultiSentiNet 方法的模态融合不充分，会丢失图片本身具有的方面信息导致准确率下降。相较于 RoBERTa-ResNet-E 方法在自建数据集上的准确率与 F_1 值分别提升了 5.86% 和 6.39%，在 MVSA-single 数据集上的准确率和 F_1 提升了 2.44% 和 1.05%；由于 RoBERTa-ResNet-E 方法提取的特征在后续的层次中进行融合和联合建模，以最终生成图像的文本描述，该方法融合层次不足，在后序融合之前由于神经网络的特征提取缺陷导致信息丢失过多，图文融合程度较低。

表 2 图文多模态对比

单位：%

Method	本文数据集		MVSA-single	
	Acc-3	F_1	Acc-3	F_1
Co-Memory+Aspect	61.98	60.41	60.57	59.71
EF-CapTrBERT	65.32	64.55	63.95	62.85
MultiSentiNet	68.86	66.61	68.34	65.46
RoBERTa-ResNet-E	69.78	67.12	69.12	68.78
本文模型	75.64	73.51	71.56	69.83

4 消融实验

为了验证模型多层次融合的性能，针对所提出的模型设置了消融实验。

LAOIS–text：表示去掉模型中的文本部分。

LAOIS–image：表示去掉模型中的图像部分。

LAOIS-1：表示去掉分类器特征融合机制获得的单层次融合情感表示。

LAOIS-2：表示去掉对比学习融合机制获得的单层次融合情感表示。

由表 3 实验结果可以看出：（1）LAOIS-text 和 LAOIS-image 的实验结果表明，单模态的图像情感分析在本文数据集上的准确率和 F_1 值为 64.95% 和 62.43%，远远低于单模态的文本情感分析。可能是因为图像中的情感信息更加难以捕捉和理解，与文本相比，表达情感的方式更为复杂。（2）LAOIS -1 去除了分类器特征融合机制，在本文数据集上的准确率和 F_1 分别下降了 9.9% 和 8.64%。LAOIS -1 仅在对比学习阶段进行交互，一定程度上解决了图文不一致性问题，但由于贴纸数据的特殊性，丢失了文本特征的重要表示导致精度大幅度的降低。（3）LAOIS-2 去除了对比学习融合机制，在本文数据集上的准确率和 F_1 分别下降了 3.76% 和 3.39%。LAOIS-2 在特征提取的后期进行融合，文本信息和图像信息的深层次融合交互解决了图文数据的关联性问题，但由于融合时序较晚，特征在神经网络中孤立传播，无法挖掘图像文本之间的局部语义关联，降低了特征间的聚合能力，无法捕获图文模态的一致性。

表 3 消融实验对比结果

单位：%

Method	本文数据集		MVSA-single	
	Acc-3	F_1	Acc-3	F_1
LAOIS-text	64.95	62.43	60.31	61.78
LAOIS-image	74.05	72.56	68.61	67.78
LAOIS-1	65.74	64.87	65.47	64.59
LAOIS-2	71.88	70.12	67.34	67.12
LAOIS	**75.64**	**73.51**	**71.56**	**69.83**

5 结语

本文提出了一个用关于多模态情感分类的社交媒体贴纸数据情感分析模型，基于多层次语义融合的方法充分挖掘贴纸数据中文本和图像模态之间存在的双向关联性。在公开数据集 MVSA-single 与自建数据集上进行实验，验证单文本模型、单图像模型、图文融合模型之间差异性，实验结果证明了本文提出模型的有效性。此外，通过消融实验，验证了多层次融合模块对模型的贡献。本文的模型仍存在许多不足之处和局限性。由于数据资源的限制，本文的实验只基于一个数据集进行，后续的工作将致力于构建一个由社交媒体产生的中文图文数据集，进一步提高模型的泛化性和鲁棒性，确保模型能够在真实世界的多样化场景中表现出色。

参考文献：

[1] 周婷 , 杨长春 . 基于多层注意力机制的图文双模态情感分析 [J]. 计算机工程与设计 , 2023, 44(6) : 1853-1859.

[2] 刘颖 , 王哲 , 房杰 , 等 . 基于图文融合的多模态舆情分析 [J]. 计算机科学与探索 , 2022,16(6):1260-1278.

[3] ZHOU L W, PALANGI H, ZHANG L, et al. Unified vision-language pre-training for image captioning and vqa[DB/OL].(2019-12-04)[2024-03-12].https://doi.org/10.48550/arXiv.1909.11059.

（下转第 196 页）

and Information Sciences (ICCAIS). Piscataway: IEEE, 2015[2024-01-19].https://ieeexplore.ieee.org/document/7338690.

[57]JOKAR P, LEUNG V C M. Intrusion detection and prevention for ZigBee-based home area networks in smart grids [J]. IEEE transactions on smart grid, 2016,9(3):1800-1811.

[58] KWON Y J, KIM H Y, LIM Y H, et al. A behavior-based intrusion detection technique for smart grid infrastructure[C/OL]//2015 IEEE Eindhoven PowerTech.Piscataway: IEEE, 2015[2024-07-01].https://ieeexplore.ieee.org/document/7232339.

[59]HONG J H, LIU C Q, GOVINDARASU M. Detection of cyber intrusions using network-based multicast messages for substation automation [C/OL]// ISGT 2014.Piscataway: IEEE, 2014[2024-02-12].https://ieeexplore.ieee.org/document/6816375.

[60]YANG Y, GAO L, YUAN Y B, et al.Intrusion detection system for IEC 61850 based smart substations [C/OL]//2016 IEEE Power and Energy Society General Meeting (PESGM). Piscataway: IEEE, 2016[2024-03-27].https://ieeexplore.ieee.org/document/7741668.

[61]YANG Y, MCLAUGHLIN K, SEZER S, et al. Intrusion detection system for network security in synchrophasor systems [C/OL]//IET International Conference on Information and Communications Technologies (IETICT 2013). London: IET, 2013[2024-06-10].https://ieeexplore.ieee.org/document/6617502.

[62]HONG J H, LIU C Q, GOVINDARASU M. Detection of cyber intrusionsusing network-based multicast messages for substation automation[C/OL]// ISGT 2014.Piscataway: IEEE, 2014[2024-04-19].https://ieeexplore.ieee.org/document/6816375.

【作者简介】

于信芳（1984—），男，辽宁辽阳人，本科，研究方向：网络信息安全、电力系统。

（收稿日期：2024-09-10）

（上接第 188 页）

[4] PENG Y X, QI J W, YUAN Y X. CM-GANs: Cross-modal generative adversarial networks for common representation learning[J]. ACM transactions on multimedia computing, communications, and applications (TOMM), 2019, 15(1): 1-24.

[5] LI C L, LI X, WANG X P, et al. FG-AGR: Fine-grained associative graph representation for facial expression recognition in the wild[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023,34(2):882-896.

[6] YOU Q, LUO J, JIN H, et al. Joint visual-textual sentiment analysis with deep neural networks[C]//Proceedings of the 23rd ACM international conference on Multimedia. New York: ACM, 2015: 1071-1074.

[7]CHEN M H, WANG S, LIANG P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning[C]// Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York: JMLR.org, 2017: 163-171.

[8] DAI W L, LIU Z H, YU T Z, et al. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition[DB/OL].(2020-10-7)[2024-02-11].https://doi.org/10.48550/arXiv.2009.09629.

[9] HU G M, ZHAO Y, LU G M,et al. Unimse: Towards unified multimodal sentiment analysis and emotion recognition[DB/OL]. (2022-11-21)[2023-12-06].https://doi.org/10.48550/arXiv.2211.11256.

[10] LL J N, LI D X, XIONG C M, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[DB/OL].(2022-02-15)[2024-03-16].https://doi.org/10.48550/arXiv.2201.12086.

[11] LIU S T, ZHANG X, YANG J F. SER30K: A large-scale dataset for sticker emotion recognition[EB/OL].(2022-10-10)[2024-01-13].https://www.semanticscholar.org/paper/SER30K%3A-A-Large-Scale-Dataset-for-Sticker-Emotion-Liu-Zhang/1d2baaf2489328baba2c4db93d44257059ded3d7.

[12] KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[DB/OL].(2021-08-05)[2023-08-16].https://doi.org/10.48550/arXiv.2108.01682.

【作者简介】

李伟（1999—），男，湖北襄阳人，硕士研究生，研究方向：多模态情感分析。

王东娟（1977—），女，陕西咸阳人，博士，副教授、硕士生导师，研究方向：信息管理与电子商务、大数据应用等。

（收稿日期：2024-09-13）