徐钦超. 基于注意力机制和对比学习的图文情感分析[J]. 智能计算机与应用,2026,16(3):172-178. DOI:10.20169/j.issn.2095-2163.24051001

基于注意力机制和对比学习的图文情感分析

徐钦超

(南京邮电大学 通信与信息工程学院, 南京 210003)

摘 要: 针对图文信息交互不充分和难以整合利用的问题,本文提出一种基于注意力机制和对比学习的图文情感分析方法。首先,对样本数据重构,并对图像进行数据增强。然后,对图文数据分别利用融合 CNN-ViT 网络和 RoBerta 网络进行特征提取,并计算图像-文本对比损失函数,完成特征对齐。随后,使用交叉注意力和多头注意力完成不同模态的特征交互和特征融合,并计算基于融合特征的对比学习损失函数。最后,通过 Softmax 分类网络对融合序列进行情感预测。在 MVSA 数据集上进行对比试验。实验结果表明,本文提出的图文情感分析方法,相比常用的情感分析方法,有着更加优异的性能表现。

关键词: 情感分析; 多模态融合; 注意力机制; RoBerta; 对比学习

中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)03-0172-07

Image and text sentiment analysis based on attention mechanism and contrastive learning

XU Qinchao

(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: This paper proposes a method for multimodal sentiment analysis based on attention mechanism and contrastive learning to address the insufficient integration and utilization of textual and visual information interaction. Firstly, the sample data is reconstructed, and data augmentation is applied to the images. Then, the textual and visual data are processed separately using a fusion of CNN-ViT network and RoBerta network for feature extraction. The image-text contrastive loss function is calculated to achieve feature alignment. Subsequently, cross-modal attention and multi-head attention are employed to accomplish feature interaction and fusion across different modalities, followed by computing the contrastive learning loss function based on the fused features. Finally, a Softmax classification network is utilized for sentiment prediction on the fused sequences. Comparative experiments on the MVSA dataset demonstrate that the proposed multimodal sentiment analysis method outperforms commonly used sentiment analysis methods in terms of performance.

Key words: sentiment analysis; multimodal fusion; attention mechanism; RoBerta; contrastive learning

0 引 言

近年来,随着社会的发展和科学的进步,社交网络已经逐渐成为人们日常交流的主要渠道。在社交网络中,每天都会有海量的用户通过各种类型的数据去表达自己的情绪和看法。这些类型的数据主要由文本、图像、视频等模态组成。如何更好地利用这些多模态数据来进行情感分析已经成为当下研究的热点之一。传统的单模态情感分析方法往往只利用文本或语音等单一模态的信息,却忽视了其他感知模态的丰富信息,因此无法全面地理解社交互动中的情感表达。而多模态情感分析则更加全面地考虑多种模态之间的联合分析,这有助于更准确地捕捉用户情感表达的多样性和丰富性。

早期的学者选择直接将不同模态的特征提取后进行拼接作为融合特征[1-2]。这种方式未能完全利用不同模态的独有特征,效果相对较差。近些年很多学者选择改进特征提取和特征融合过程来提升情感分析的效果。Deshmukh 等学者[3]为了最大化不同模态数据之间的相似性,选择将多模态数据进行特征学习并映射至相关性度量空间进行计算。Zhang 等学者[4]选择将模态数据发送至内部类映射

作者简介: 徐钦超(1998—),男,硕士研究生,主要研究方向:情感分析。Email:molincx@163.com。

收稿日期: 2024-05-10

模型中，通过计算最大平均差异获得标签，然后通过类感知注意力门控递归单元进行标签构建。孙文飞等学者[5]选择引入注意力机制对特征进行提取，并基于张量进行特征融合。Wang 等学者[6]提出一种两阶段学习框架 AMSA，自适应地学习模态的相关性和互补性进行动态融合。张晋敏等学者[7]选择利用融合注意力机制的 BiGRU 模型进行情感融合分类。王顺杰等学者[8]提出使用正交约束的自注意力机制生成各个模态的语义图，再通过图卷积获得含有方面词的语义图表示。Tong 等学者[9]提出一种双向递归模型 DRMM，利用交替双重注意力选择信息特征进行互补。

多模态情感分析的关键是对不同模态信息的融合利用，而这些不同模态数据仍会存在交互不充分和难以整合利用的问题。为了解决上述问题，本文提出一种基于注意力机制和对比学习的图文情感分析方法。使用注意力机制对图文特征进行提取、交互和融合，除此之外构建基于融合特征的对比学习任务，优化情感分析模型的效果。

1　情感分析框架

本文提出的图文情感分析模型，其整体结构如图 1 所示。该模型首先对数据样本中的文本进行拼接重构，对图像进行线性插值重构，并进行数据增强；然后，对于文本特征使用 Roberta 模型进行特征提取，对于图像数据先用 ResNet-50 网络提取局部特征，将其处理后送入 ViT 网络中进行全局特征构建，并对图像-文本特征进行特征对齐计算。随后，使用交叉注意力和多头注意力进行模态特征交互和融合，得到图文融合序列。接着，根据数据增强前后的图文融合序列构建对比学习任务。最后，将图文融合序列送入 Softmax 分类器预测情感极性。

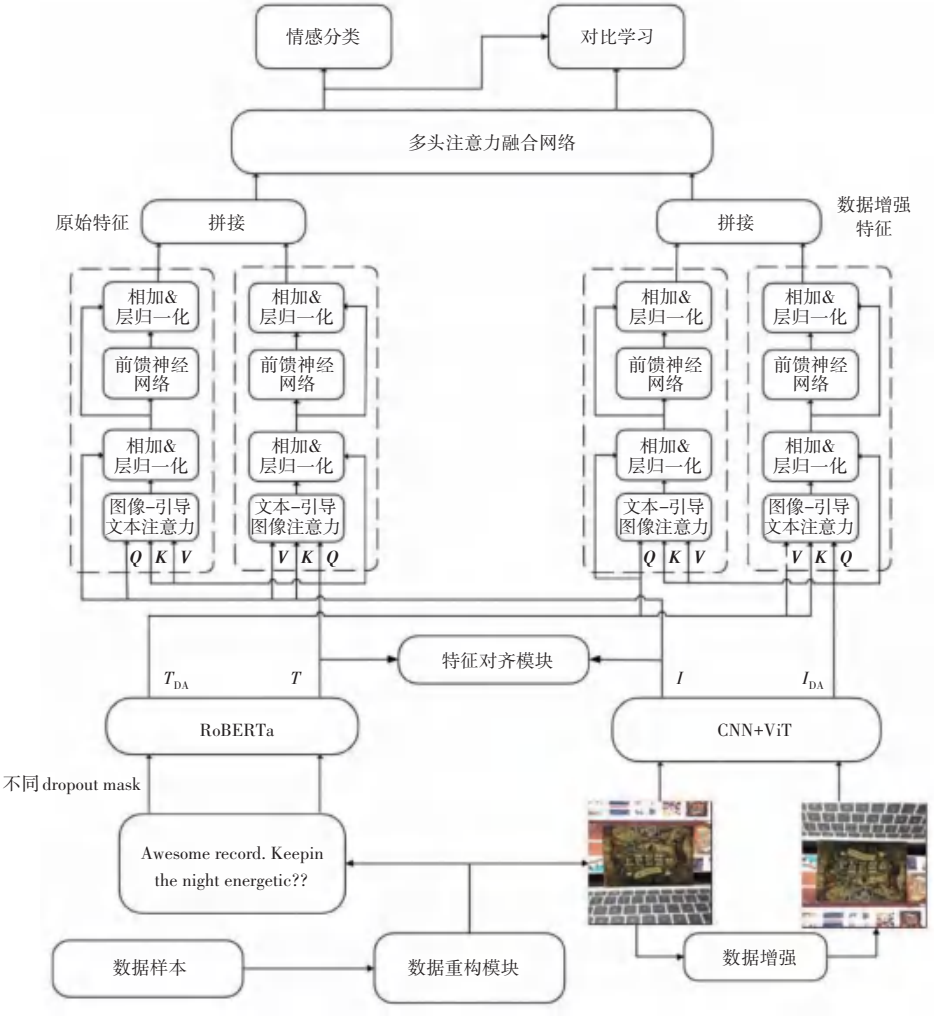

图 1　基于注意力机制和对比学习的图文情感分析模型

Fig. 1　Image-text sentiment analysis model based on attention mechanism and contrastive learning

1.1 图文特征编码网络

1.1.1 文本特征提取

本文选择使用 RoBerta 模型[10]对文本进行特征提取。RoBerta 模型是对 Bert 模型[11]训练优化得到的预训练模型。首先对文本进行预处理,去除其中无意义的特殊符号,然后使用 BBPE 分词器进行分词。将得到的序列进行处理并映射为输入序列 $\boldsymbol{E}_{\text{input}}=\{E_c,E_1,E_2,\cdots,E_s\}$, $E\in\mathbb{R}^{n_t\times d_t}$, 其中步骤包含添加分类用标记[cls]、分句标记[sep]、位置编码等。再将输入序列送入 Transformer 编码器中进行特征编码,得到的输出为所提取到的文本特征 $\boldsymbol{T}$,对此表示为:

$$\boldsymbol{T}=\text{TE}(\boldsymbol{E}_{\text{input}})=\{t_c,\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_s\} \tag{1}$$

其中, $\boldsymbol{T}\in\mathbb{R}^{n_t\times d_t}$, TE(·)表示 Transformer 编码器结构。

1.1.2 图像特征提取

本文采用融合 CNN 和 ViT[12]的网络模型对图像进行特征提取。其具体结构如图 2 所示。

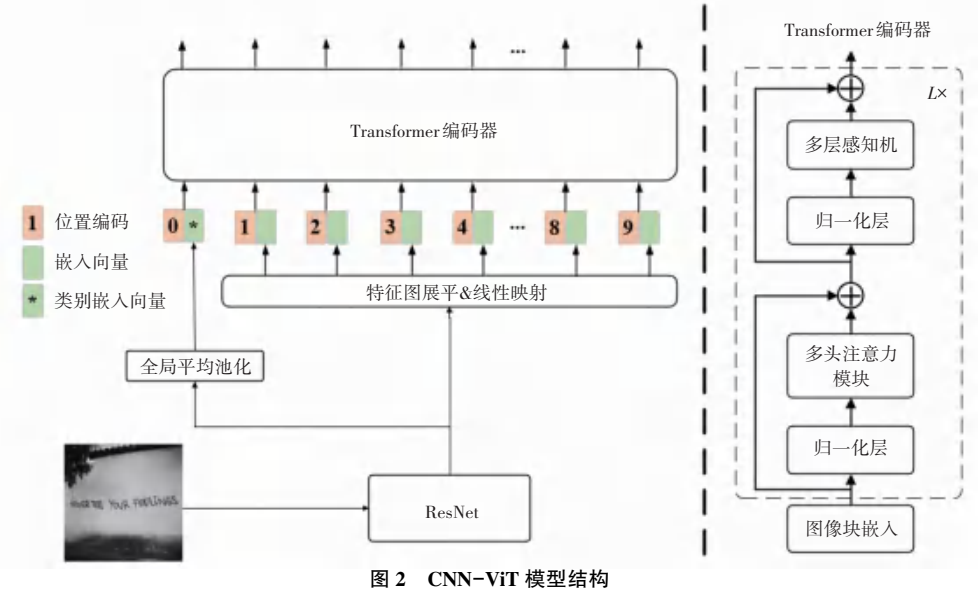

图 2 CNN-ViT 模型结构

Fig. 2 Structure of CNN-ViT

本文选择使用 ResNet-50[13]作为第一步提取特征的卷积神经网络。将图片转化为 RGB 三通道向量送入 ResNet-50 网络中提取到视觉特征 $\boldsymbol{I}_f$, 选择 ResNet 最后一个卷积块的输出作为特征图,研究推得的数学公式为:

$$\boldsymbol{I}_f=\text{ResNet}(\boldsymbol{I}_{\text{input}}) \tag{2}$$

其中, $\boldsymbol{I}_f\in\mathbb{R}^{p\times p\times d_i}$。这里,$p$ 表示通过 ResNet 的特征图高度与宽度, d_i 表示特征图的嵌入维度。

此后,调整特征图尺寸,将其线性变化和展平,得到图像的局部特征表示 $\boldsymbol{I}_{\text{local}}$, 用到的公式为:

$$\boldsymbol{I}_{\text{local}}=\text{flatten}(\boldsymbol{I}_f\boldsymbol{W}_I+\boldsymbol{b}_I) \tag{3}$$

其中, $\boldsymbol{I}_{\text{local}}=\{i_1,i_2,\cdots,i_{n_i'}\}$; $\boldsymbol{I}_{\text{local}}\in\mathbb{R}^{n_i'\times d_i}$, $n_i'=p\times p$; flatten(·)表示展平函数; $\boldsymbol{W}_I$ 和 $\boldsymbol{b}_I$ 表示线性变换的权重矩阵和偏置矩阵

再对特征图进行全局平均池化,得到代表图像全局特征表示 $\boldsymbol{I}_{\text{global}}$, 并作为[cls]标记。具体公式如下:

$$\boldsymbol{I}_{\text{global}}=\text{AdaptiveAvgPool2d}(\boldsymbol{I}_f\boldsymbol{W}_I+\boldsymbol{b}_I) \tag{4}$$

其中, $\boldsymbol{I}_{\text{global}}\in\mathbb{R}^{1\times d_t}$, AdaptiveAvgPool2d(·)表示全局二维平均池化操作。这里, $d_t=768$。

随后,将图像的全局特征表示和局部特征表示拼接,添加位置编码之后送入 Transformer 编码器中,得到最后的编码器输出图像特征序列, 用如下公式进行定义:

$$\boldsymbol{I}_{\text{embed}}=\text{Add}(\text{Concat}(\boldsymbol{I}_{\text{global}},\boldsymbol{I}_{\text{local}}),\boldsymbol{I}_{\text{pos}}) \tag{5}$$

$$\boldsymbol{I}_{\text{encoded}}=\text{TE}(\boldsymbol{I}_{\text{embed}})=\{i_{\text{CLS}},i_1,i_2,\cdots,i_{n_i'}\} \tag{6}$$

其中,Concat(·)表示拼接操作;Add(·)表示加运算; $\boldsymbol{I}_{\text{pos}}$ 表示位置编码; TE(·) 表示 Transformer 编码器。

通过上述步骤对图像数据进行特征提取。ResNet-50 网络提取到的局部特征可以为 ViT 提供更加精细化的信息。而 ViT 的注意力机制可以为全局特征和局部特征进行自适应权重配比。两者结合可以提取到更加完整的图像情感信息。

1.1.3 特征对齐模块

本文对图像特征和文本特征进行计算图像-文本对比损失(Image-Text Contrastive Loss, ITC)[14], 完成特征对齐模块的设计。具体为分别计算文本到

图像的相似度（L_{t2i}）和图像到文本的相似度（L_{i2t}），最后将其平均。对此过程可以表示为：

$$L_{i2t} = -\frac{1}{|I|}\sum_{i\in I}\log\left[\frac{\exp(\mathrm{sim}(i,t^{+})/\tau)}{\sum_{t'\in T}\exp(\mathrm{sim}(i,t')/\tau)}\right] \tag{7}$$

$$L_{t2i} = -\frac{1}{|T|}\sum_{t\in T}\log\left[\frac{\exp(\mathrm{sim}(t,i^{+})/\tau)}{\sum_{i'\in I}\exp(\mathrm{sim}(t,i')/\tau)}\right] \tag{8}$$

$$L_{\mathrm{ITC}} = \frac{L_{i2t}+L_{t2i}}{2} \tag{9}$$

其中，sim(·)表示计算相似度的函数，本处使用余弦相似度计算；τ 表示对比学习温度系数；(i, t^{+}) 表示图像特征 i 和与其正对的文本特征 t^{+}；(t, i^{+}) 表示文本特征 t 和与其正对的图像特征 i^{+}，通过这种方式体现上述过程中真实独热相似度的计算过程。

1.2　图文特征交互与融合网络

模态交互网络主要用于实现图像和文本的交互，其实现原理基于交叉注意力。以提取文本引导的图像特征为例介绍模态交互过程。选择使用特征提取后的文本特征（$\boldsymbol{T}$）计算查询向量（$\boldsymbol{Q}$），使用图像特征 $\boldsymbol{I}$ 计算键向量（$\boldsymbol{K}$）和值向量（$\boldsymbol{V}$），计算公式具体如下：

$$\boldsymbol{Q}_i = \boldsymbol{W}_i^{Q}\boldsymbol{T} \tag{10}$$

$$\boldsymbol{K}_i = \boldsymbol{W}_i^{K}\boldsymbol{I} \tag{11}$$

$$\boldsymbol{V}_i = \boldsymbol{W}_i^{V}\boldsymbol{I} \tag{12}$$

然后，将其送入多头注意力机制模块中进行文本引导图像特征交互计算。在通过多个单独注意力对其进行运算之后，将其进行多头结果拼接和线性变化，进而得到模块输出文本引导图像特征 $\boldsymbol{I}^{\mathrm{TG}} = \{i_{\mathrm{cls}}^{\mathrm{TG}}, i_1^{\mathrm{TG}}, \cdots, i_N^{\mathrm{TG}}\}$，将文本引导的图像特征交互模块记为 $\mathrm{TF}^{\mathrm{TGI}}(\boldsymbol{T},\boldsymbol{I})$。

同理，将图像引导结合文本特征交互模块记为 $\mathrm{TF}^{\mathrm{IGT}}(\boldsymbol{I},\boldsymbol{T})$，最终得到的图像引导的文本特征为 $\boldsymbol{T}^{\mathrm{IG}} = \{t_{\mathrm{cls}}^{\mathrm{IG}}, t_1^{\mathrm{IG}}, \cdots, t_N^{\mathrm{IG}}\}$，进一步推得：

$$\boldsymbol{T}^{\mathrm{IG}} = \mathrm{TF}^{\mathrm{IGT}}(\boldsymbol{I},\boldsymbol{T}) \tag{13}$$

其中，通过图像特征计算查询向量（$\boldsymbol{Q}$），通过文本特征计算键向量（$\boldsymbol{K}$）和值向量（$\boldsymbol{V}$）。

之后将文本引导图像特征 $\boldsymbol{I}^{\mathrm{TG}}$ 和图像引导文本特征 $\boldsymbol{T}^{\mathrm{IG}}$ 进行拼接，组成多模态特征序列。再将拼接序列送入 Transformer 编码器中进行基于特征级别的文本图像融合，得到融合之后的多模态特征向量 $\boldsymbol{F}$。具体计算过程如下：

$$\boldsymbol{F} = \mathrm{TE}(\mathrm{concat}(\boldsymbol{T}^{\mathrm{IG}}, \boldsymbol{I}^{\mathrm{TG}})) = \{f_1, f_2, \cdots, f_N\} \tag{14}$$

融合后的多模态特征向量无法直接进行情感分类任务。所以本文使用一个注意力层来获取后续用于情感分类和对比学习任务的图文融合表示，计算公式具体如下：

$$a_i' = \mathrm{Relu}(f_i\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{15}$$

$$a_i = \exp\left(\frac{a_i'}{\sum_{j=1}^{N} a_j'}\right) \tag{16}$$

$$R = \sum_{i=1}^{N} a_i f_i \tag{17}$$

其中，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 表示线性变化权重矩阵；$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 表示对应项的偏置；Relu(·)表示激活函数。

1.3　对比学习任务

1.3.1　数据集重构

研究中，需要对使用数据集的部分数据进行重构，具体重构过程如图 3 所示。首先对于选择的图像-文本对，将图像通过线性插值的方式对图像数据进行第一次数据增强；而对于文本数据，本节选择直接将 2 个样本对中的文本数据进行拼接，并替换原本样本。通过上述步骤对数据集进行重构，计算过程如下：

$$I_k = \lambda \cdot I_i + (1-\lambda)\cdot I_j \tag{18}$$

$$T_k = \mathrm{concat}(T_i, T_j) \tag{19}$$

图 3　图像-文本重构过程

Fig. 3　Image-text reconstruction process

1.3.2 对比学习

首先对于重构后数据集中的图像样本,将其进行 RandAugment 数据增强。并且考虑到颜色改变对图像特征带来的差异,选择将 RandAugment 策略中与颜色改变相关的方法删去。而对于文本将不进行额外的数据增强,而是利用 SimCSE[15] 中使用的方法,即将文本进行复制,并将 2 份文本输入至文本提取模型中,利用 Roberta 中内置的 dropout mask 机制得到 2 个相似的文本特征向量构成文本正负样本对。将上述数据样本送入前文构建的特征编码、交互、融合网络得到数据增强前后的多模态特征表示 R 和 R_{DA},计算对比学习目标损失,此处使用的是修改 SimCSE 的目标损失函数。前文提出的交互和融合网络定义为 MF(·),数据增强模块定义为 DA(·),数据集重构步骤定义为 ReBuild(·),计算流程如下所示:

$$(T^{'},I^{'}) = \mathrm{ReBuild}(T,I) \tag{20}$$

$$R = \mathrm{MF}(T^{'},I^{'}) \tag{21}$$

$$R_{\mathrm{DA}} = \mathrm{MF}(T^{'},\mathrm{DA}(I^{'})) \tag{22}$$

$$L_{\mathrm{cl}} = \mathrm{infoNCE}(R,R_{\mathrm{DA}}) = -\log \frac{e^{\frac{\mathrm{sim}(R_i,R^{'}_{\mathrm{DA}i})}{\gamma}}}{e^{\frac{\mathrm{sim}(R_i,R^{'}_{\mathrm{DA}i})}{\gamma}} + m\sum_{j=1,j\neq i}^{N} e^{\frac{\mathrm{sim}(R_i,R_{\mathrm{DA}j})}{\gamma}}} \tag{23}$$

其中,sim(·)表示相似度计算。

在计算正对的相似度时,使用 $(R_i,R^{'}_{\mathrm{DA}i})$,即同一文本通过不同 dropout mask 机制得到正样本对后处理得到的融合特征。计算负对的相似度时,使用 $(R_i,R_{\mathrm{DA}j})$,即不考虑 dropout mask 机制得到的融合特征。

1.4 Softmax 分类器

Softmax 函数是一种常用的归一化函数,将一个包含多个任意实数的 K 维向量映射为一个 K 维的概率分布,即将每个元素映射至 0 到 1 的区间内,并使所有元素的总和为 1。这样在处理多分类任务时,K 维的概率分布就代表了 K 个类别的概率情况。假设 Softmax 函数的输入是一个包含 K 个元素的向量 $\vec{z} = [z_0,z_1,\cdots,z_K]$,具体计算公式如下:

$$P = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{24}$$

最终,得到该样本针对不同类别的概率向量,选择概率最大的类别作为该样本的预测类别。

1.5 损失函数

本文将图文对比损失(L_{ITC})、对比学习任务损失(L_{cl})、预测分类损失(L_{sc})进行加权和作为最后的损失函数 Loss 优化整个情感分析模型,具体公式如下:

$$\mathrm{Loss} = L_{\mathrm{sc}} + \lambda_{\mathrm{ITC}} L_{\mathrm{ITC}} + \lambda_{\mathrm{cl}} L_{\mathrm{cl}} \tag{25}$$

预测分类损失 L_{sc} 为交叉熵损失,计算预测标签与样本真实标签之间的差异。

2 实验研究

2.1 数据集

本文选用的数据集为 MVSA 数据集[16],其样本数据为从 Twitter 上收集的图文评论,其情感标签分为积极、中性、消极。MVSA 数据集包含 2 部分,一部分为 MVSA-Single、包含 5 129 个图文对,其样本标签由一位工作人员进行标注,每个样本仅包含一组模态情感标签;另一部分为 MVSA-Multiple、包含 19 600 个图文对,其样本标签由 3 位工作人员进行标注,故每个样本包含 3 组情感标签。使用前需要对数据集进行预处理。对 MVSA-Single 数据集,需要预先删除其中情感极性相反的图文对。并且,如果图文对中有一模态情感标签为中性,则将另一模态消极或积极的情感标签作为此图像对的修正情感标签;而对 MVSA-Multiple 数据集,统计 3 组标签中出现次数最多的情感极性作为当前图文对的情感标签,如果出现次数相等,则删除此图文对。通过上述处理之后,MVSA-Single 数据集使用的图文对共计 4 511 对,MVSA-Multiple 数据集使用的图文对共计 17 024 个图文对。本文选择将 2 个数据集根据 8∶1∶1 的比例划分为训练集、验证集和测试集,具体见表 1。

2.2 参数设置

本实验基于 Python3.8 构建模型,利用 Nvidia RTX3080Ti GPU 进行模型的训练。针对 MVSA-Single 和 MVSA-Multiple 设置的 batchsize 分别为 32 和 16。采用 Adamw 优化器,学习率设为 2e-5。训练的 epoch 设置为 100,Dropout 设置为 0.2。λ_{ITC} 和 λ_{cl} 都设置为 0.9。

2.3 实验结果与分析

为验证本文提出的图文融合情感分析模型的有效性,将其与一些基线模型在不同数据集上进行性能比较,具体的评测指标选择准确率(Acc)和 Weighted-$F1$ 分数。实验结果见表 2。

表 1　数据集详细信息

Table 1　Dataset details

数据集	标签	训练集	验证集	测试集	总计
MVSA-Single	积极	2 145	269	269	2 683
	中性	376	47	47	470
	消极	1 086	136	136	1 358
	总计	3 607	452	452	4 511
MVSA-Multiple	积极	9 054	1 132	1 132	11 318
	中性	3 526	441	441	4 408
	消极	1 038	130	130	1 298
	总计	13 618	1 703	1 703	17 024

表 2　情感识别结果

Table 2　Sentiment recognition results

类别	对比模型	MVSA-Single		MVSA-Multiple	
		Acc	Weighted - $F1$	Acc	Weighted - $F1$
文本	BiLSTM[17]	0.701 2	0.650 6	0.679 0	0.679 0
	BERT[7]	0.711 1	0.697 0	0.675 9	0.662 4
图像	ResNet-50[13]	0.646 7	0.615 5	0.618 8	0.609 8
	OSDA[18]	0.667 5	0.665 1	0.666 2	0.662 3
多模态	MultiSentiNet[19]	0.698 4	0.698 4	0.688 6	0.681 1
	CFF-ATT[20]	0.714 4	0.710 6	0.696 2	0.693 5
	MVAN-M[18]	0.729 8	0.729 8	0.723 6	**0.723 0**
	MGNNS[21]	0.737 7	0.727 0	0.724 9	0.693 4
	本文模型	**0.754 4**	**0.734 7**	**0.730 4**	0.715 1

通过表 2 分析得知，相比单模态的情感分析方法，多模态的分析方法能够获取到来自不同模态的特征，效果更加优异。而本文提出的基于注意力机制和对比学习的情感分析模型，在 MVSA-Single 上 Acc 为 0.754 4，Weighted - $F1$ 分数为 0.734 7；在 MVSA-Multiple 上 Acc 为 0.730 4，Weighted - $F1$ 分数为 0.715 1。相比现阶段常用的情感分析方法，其情感识别效果更好。

而本模型在 MVSA-Multiple 上相比 MVAN-M 模型有一定的差距，相差 0.007 9。通过使用不断更迭的记忆网络进行特征学习，得到了更好的特征表示。但是在其余几项指标表现中，本文提出的模型效果远超其实验结果。这证明本文的模型结构在一定程度上取得了优势，并且适用范围更广。而相比其余的多模态情感分类模型，本文通过基于注意力机制和对比学习的设计，获得了更好的模态情感表示，并在此基础上进行交互融合，使模型对于不同情感极性的特征更为敏感，得到了相对较好的情感识别效果。

2.4　消融实验与分析

为验证不同模块对整体模型的影响，实验还测试了本文各模块的性能对比。消融实验见表 3。

表 3　消融实验

Table 3　Ablation experiment

模型	MVSA-Single		MVSA-Multiple	
	Acc	Weighted - $F1$	Acc	Weighted - $F1$
CNN-ViT+Roberta	0.714 6	0.703 4	0.685 2	0.673 4
+模态交互、融合	0.734 5	0.712 3	0.717 6	0.701 6
+CL	0.727 8	0.706 8	0.702 3	0.690 3
本文模型	**0.754 4**	**0.734 7**	**0.730 4**	**0.715 1**

表 3 中,CNN-ViT+RoBERTa 代表对样本进行特征提取后直接拼接送入分类器的实验结果;“+模态交互、融合”指的是,在特征提取网络之后使用注意力机制进行跨模态交互和模态融合后的分类结果;“+CL”指的是在 CNN-ViT+RoBERTa 情况下,添加单独模态的对比学习任务,并修改实验目标函数之后的分类结果。

通过表 3 的结果可以看出,在特征提取网络不变的情况下,添加特征交互和注意力机制融合后,得到的特征序列学习了更多的模态信息,取得了最大的效果提升。因为特征交互和特征融合模块能帮助 2 个不同的特征进行互相学习和利用,帮助情感类别的预测。而单独添加对比学习任务,也取得了一定的效果。不过由于选择的模态融合方法为直接拼接、而不是通过注意力机制融合,取得的效果相对比较微弱。所以本文选择将特征交互、融合以及对比学习进行串联使用,由此得到了最显著的效果提升,同时也证明本文提出的模型可行性和有效性,其具有一定的实用价值和研究价值。

3 结束语

本文从社交媒体中的图像和文本数据出发,设计出一种基于注意力机制和对比学习的图文情感分析模型。首先,对数据集中的样本进行重构处理,并进行数据增强。然后,使用 Roberta 网络提取文本特征,使用融合 CNN-ViT 网络提取图像特征,并进行特征对齐。接着,利用交叉注意力和多头注意力完成对文本、图像的特征交互和融合。随后,基于数据增强前后的图文融合序列,构建对比学习任务。最后,使用图文融合序列进行情感分类预测。实验结果表明,该模型能够更加有效地利用来自不同模态的信息进行情感分类。

参考文献

[1] SOLEYMANI M, GARCIA D, JOU B, et al. A survey of multimodal sentiment analysis[J]. Image and Vision Computing, 2017, 65: 3-14.

[2] 李勇敢, 周学广, 孙艳, 等. 中文微博情感分析研究与实现[J]. 软件学报, 2017, 28(12):3183-3205.

[3] DESHMUKH S, ABHYANKAR A, KELKAR S. DCCA and DMCCA framework for multimodal biometric system [J]. Multimedia Tools and Applications, 2022, 81(17): 24477-24491.

[4] ZHANG Ke, ZHU Yunwen, ZHANG Wenjun, et al. Transfer correlation between textual content to images for sentiment analysis[J]. IEEE Access, 2020, 8: 35276-35289.

[5] 孙文飞, 张云华. 基于多模态的图文情感分析[J]. 智能计算机与应用, 2023, 13(12):102-106.

[6] WANG Jingyao, MOU Luntian, MA Lei, et al. AMSA: adaptive multimodal learning for sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(3s): 1-21.

[7] 张晋敏, 李旭芳, 樊弟军. 基于 BiGRU 模型的多模态网络舆情情感分析[J]. 智能计算机与应用, 2024, 14(1):191-193.

[8] 王顺杰, 蔡国永, 吕光瑞, 等. 方面级多模态协同注意图卷积情感分析模型[J]. 中国图象图形学报, 2023(12):3838-3854.

[9] TONG Meihan, WANG Shuai, CAO Yixin, et al. Image enhanced event detection in news articles[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9040-9047.

[10] LIU Yinhan, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv, 1907. 11692, 2019.

[11] DEVLIN J, CHANG Mingwei, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv, 1810. 04805, 2018.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv, 2010. 11929, 2020.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE, 2016: 770-778.

[14] LI Junnan, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation [J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.

[15] GAO Tianyu, YAO Xingcheng, CHEN Danqi. SIMCSE: Simple contrastive learning of sentence embeddings[J]. arXiv preprint arXiv, 2104. 08821, 2021.

[16] NIU Teng, ZHU Shiai, PANG Lei, et al. Sentiment analysis on multi-view social data[M]//TIAN Q, SEBE N, QI G J, et al. Multi Media Modeling. Lecture Notes in Computer Science(). Cham: Springer, 2016, 9517: 15-27.

[17] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv, 1508. 01991, 2015.

[18] YANG Xiaocui, FENG Shi, WANG Daling, et al. Image-text multimodal emotion classification via multi-view attentional network [J]. IEEE Transactions on Multimedia, 2020, 23: 4014-4026.

[19] XU Nan, MAO Wenji. Multisentinet: A deep semantic network for multimodal sentiment analysis[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York:ACM, 2017: 2399-2402.

[20] ZHANG Kang, GENG Yushui, ZHAO Jing, et al. Sentiment analysis of social media via multimodal feature fusion [J]. Symmetry, 2020, 12(12): 2010.

[21] YANG Xiaocui, FENG Shi, ZHANG Yifei, et al. Multimodal sentiment detection based on multi-channel graph neural networks [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021,1: 328-339.