计算机科学
COMPUTER SCIENCE

基于多视角的图像文本情感分析

高玮军, 孙子博, 刘书君

引用本文

高玮军, 孙子博, 刘书君. 基于多视角的图像文本情感分析[J]. 计算机科学, 2024, 51(11A): 231200163-8.

GAO Weijun, SUN Zibi, LIU Shujun. Sentiment Analysis of Image-Text Based on Multiple Perspectives [J]. Computer Science, 2024, 51(11A): 231200163-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200163

基于多视角的图像文本情感分析

高玮军　孙子博　刘书君
兰州理工大学计算机与通信学院　兰州 730000
(gaoweijun@lut.edu.cn)

摘　要　在社交媒体中,人们往往首先被图片中的人物表情所吸引,直接触及到情感。然而,对于情感的完整表达,场景也扮演着不可或缺的角色,为情感分析提供了必要的背景和支持。但许多学者忽视了场景在情感表达中的重要性,导致结果并非最优。针对图文双模态情感分析模型存在忽略多模态间的对齐、图片特征提取不充分和模型泛化能力不高的问题,提出了多视角图像文本情感分析网络(Multi-view Image-Text Emotion Analysis Network Model,MITN)。在图像特征提取中,在面部表情方面加入注意力机制来更好地捕捉人物面部表情,在场景方面加入空洞卷积引入膨胀率来增大感受野,并利用 Places 数据集对 Scene-VGG 进行迁移学习训练,以此来充分利用场景。使用 BERT+BiGRU 来提取文本表达特征,在多模态情感数据集 MVSA 上的实验验证了所提模型的有效性。

关键词:多模态;情感分析;多视角;迁移学习;注意力机制
中图分类号　TP391

Sentiment Analysis of Image-Text Based on Multiple Perspectives

GAO Weijun,SUN Zibi and LIU Shujun
School of Computer and Communication,Lanzhou University of Technology,Lanzhou 730000,China

Abstract　In the realm of social media,facial expressions of characters in pictures often captivate our attention first,directly evoking strong emotional responses. However,for a truly comprehensive emotional expression,scenes play a pivotal role,serving as a crucial backdrop and support for emotional analysis. Scenes provide context,setting the tone and atmosphere for the emotions being expressed. Regrettably,numerous scholars have failed to fully recognize the significance of scenes in emotional expression,often focusing solely on facial expressions. This oversight has led to suboptimal outcomes in sentiment analysis,missing out on the rich emotional nuances that scenes can provide. To address these challenges,we propose the multi-view image text sentiment analysis network(MITN). This innovative approach takes into account both facial expressions and scenes,providing a more comprehensive analysis of emotional expression. In MITN,we enhance image feature extraction by incorporating an attention mechanism that meticulously captures the facial expressions of characters. At the same time,dilated convolution is introduced to broaden the receptive field,focusing on the intricate details of the scene. Moreover,we leverage the Places dataset for transfer learning training of Scene-VGG. This allows us to fully utilize the vast amount of scene information available,enhancing the accuracy and depth of our emotional analysis. The effectiveness of MITN is rigorously tested through experiments on the multimodal sentiment dataset MVSA. Utilizing BERT+BiGRU to extract text expression features,our model demonstrates superior performance in sentiment analysis,accurately capturing the emotional nuances present in both facial expressions and scenes. This comprehensive approach offers a new perspective in sentiment analysis,paving the way for more accurate and nuanced understanding of emotional expression in social media.

Keywords　Multi-modal,Sentiment analysis,Multi-view,Transfer learning,Attention mechanism

情感是人们内心对外界事物所持肯定或否定态度的心理体现,它在人们的交流、学习和决策过程中起着至关重要的作用。随着社交媒体和在线平台的快速发展,情感分析已成为一种新兴技术,用于解析带有情感色彩的各种模态信息,并预测用户的行为。从早期的文本评论分析,到现在的多模态情感分析,情感分析技术经历了显著的进步。

早期的情感分析主要聚焦于文本内容,通过文本中的词句来推断用户的情绪状态[1-4]。然而,随着社交媒体中视频、图片以及图片配文的大量涌现,单一的文本情感分析已不能满足需求。图像中的场景、颜色、面部表情等视觉元素同样承载着丰富的情感信息。因此,近年来,多模态情感分析[5-10]逐渐成为研究热点,旨在结合文本和图像等多种模态信息,以更全面地捕捉和解析情感。

尽管多模态情感分析取得了显著的进展,但仍存在一些亟待解决的问题。首先,当前的图像特征提取方法往往只关注表面的视觉信息,而忽略了深层的语义特征。例如,在一张

基金项目:国家自然科学基金(51668043)
This work was supported by the National Natural Science Foundation of China(51668043).
通信作者:孙子博(941909610@qq.com)

描绘日落场景的图片中，除了基本的颜色、形状等视觉特征外，还蕴含着宁静、平和等深层的情感语义。这些语义特征对于准确理解图片中的情感至关重要。然而，许多现有的模型并未全面考虑这些深层语义特征，导致情感分析的结果不够准确。其次，模态之间的对齐问题也是多模态情感分析面临的一大挑战。在多模态情感分析中，文本和图像等不同模态的信息需要进行有效的对齐和融合，以捕捉跨模态的情感联系。然而，由于不同模态之间的数据差异和语义鸿沟，实现有效的模态对齐仍然是一个具有挑战性的问题[1-4,6-7]。这导致了模型在解析复杂情感时的困难，例如当文本和图像表达的情感不一致或相互矛盾时。

为了解决上述问题，本文提出了一种多视角图文情感分析网络（Multi-view Image-Text Emotion Analysis Network model，MITN）。该模型通过场景迁移学习获取场景特征捕捉能力，对图片进行多视角的场景和面部表情分析，以获得更全面的图像特征；同时，采用 BERT+Bi-GRU 结构处理文本，以更好地理解文本语义信息。此外，模型还实现了图像和文本之间的跨模态注意力对齐，使得不同模态的信息可以更加充分地交互和融合。最后，通过多层感知器和堆叠池模块构建了多模态融合模块，在融合过程中降低噪声，提高情感分析的准确性。

1 相关工作

1.1 图像情感分析

图像情感分析的基本任务是对输入的图片进行情感极性分类[11]，包括积极、消极或中性。以往研究者对图像进行情感分析时，通过基于图像情感感知检索的方法来寻找情感和图像视觉特征之间的联系[12-13]。Borth 等[14]构建了一个由 1 200 个形容词名词对（Adjective Noun Pair，ANP）组成的大型视觉情感库，同时在该库的基础上分别提出了情感银行（Sentiment Bank）和视觉情感题库（Visual Sentiment Ontology，VSO）的情感探测器来提取输入图像中的层表示。You 等[15]研究了局部图像区域对视觉情感分析的影响，通过注意力机制来共同发现相关的局部区域，并在这些局部区域之上构建情感分类器。Simonyan 等[16]为提高大规模图像识别（ImageNet）的分类精度，提出了名为 VGG-* 的超深卷积网络，尽管 VGG 模型的主要应用在于物体识别，但其强大的特征提取能力使其在图像情感分析中也具有潜在的应用价值。VGG 模型能够学习图像中的层次化特征，这些特征可以被用于捕捉与情感相关的视觉模式。Zhou 等[17]描述了 Places 数据库和相应的 Places-CNNs 模型，该模型是为场景分类任务而设计的。Places 数据库包含了丰富多样的场景图像，这些图像涵盖了各种环境和情感氛围。因此，Places-CNNs 模型学到的特征可能包含与场景情感相关的信息。通过将这些特征用于图像情感分析，我们可以期望捕捉到与场景相关的情感表达。

1.2 文本情感分析

文本情感分析指根据一段文字、句子、词语中包含的情感信息来识别文本中包含的情感类别。文本情感分析已经在不同类型的数据集上被广泛地研究。如在推特数据集中，Muhammad 等[18]提出了一个使用混合分类方案对推文进行分类的统一框架，以此来提高基于 Twitter 的情感分析系统的性能。Hamouda 等[19]利用一个名为“情感网络”的情感词汇，通过给每个单词分配积极、中性和消极的情感得分来对评论进行分类。Tang 等[20]将特定句子的词语嵌入（SSWE）特征与手工制作的特征相结合建立了监督学习框架，该框架用于消息级 Twitter 情感分类的深度学习系统。Yang 等[21]利用注意机制帮助网络选择重要的单词和句子，开发了一种用于文档级情绪分类任务的层次注意网络（HAN）。在新闻数据集中，Chiong 等[22]提出了一种基于情绪分析的方法，该方法通过披露新闻对金融市场进行预测，获得了出色的性能。

1.3 多模态情感分析

在文本基础上，加入图片信息能提供更生动的描述，传达更准确和丰富的情感信息，展现文本可能隐藏的信息。通过结合文本和图片信息，不仅可以提供更生动的描述，而且有助于深入挖掘文本可能隐藏的信息，这使得图文数据库在研究多模态相关性时变得尤为关键。图文数据库与传统的图像数据库不同，社交图像与文字通常相互关联，这使得情感分析变得非常重要，并且也有越来越多的研究人员研究多模态间的相关性。Xu 等[23]为了解决两种模态之间复杂的相关性而提出了一种双向多级注意（BDMLA）模型。Huang 等[24]利用判别特征以及视觉和语义内容之间的内部相关性，提出了图像-文本情感分析模型，即深度多模态注意力融合（DMAF），用于情感分析的混合融合框架。Xu 等[25]发现了多模态学习方法捕获图像和文本之间的关系仅停留在区域级别，忽略了通道和语义密切相关的事实，因此提出了一种渐进式双注意力模块来捕获图像和文本之间的相关性，从内容信息的角度学习图像-文本的联合表示。Yang 等[26]提出了 ConFEDE 框架，结合对比学习和特征分解，以文本为中心学习多模态信息的表示，提升视频情绪分析的准确性，并在多个数据集上优于基线方法。Fan 等[27]发现现有方法大多采取粗粒度注意力机制，如果方面级有多个单词或更大的上下文，则可能会带来信息丢失，因此提出了一种新颖的多粒度注意力网络（MGAN）模型，用于方面级别的情感分类。Zhang 等[28]提出了一个面向误差的混合课程学习框架，包括对话水平和话语水平的课程。通过构建难度测量器和情感相似性策略，提高了对话中的情绪识别模型的性能。

由于多模态融合在多模态情绪分析中起着重要的作用，因此有学者提出了关于设计不同模态之间的融合策略的工作。Poria 等[29]提出了一种基于注意机制的融合网络（ATFusion）来融合不同模式的特征。Zadeh 等[30]开发了一种新的融合方法，称为张量融合，用于多模态情绪分析，该方法明确地建模了单峰、双峰和三峰动力学。门控多模态单元（GMU）模型由 Arevalo 等[31]提出，通过学习控制输入模式对单元激活的影响程度来进行数据融合。Liu 等[32]提出了低秩多模态融合，一种利用低秩张量进行的低秩多模态融合方法，并证明了其在情绪分析任务上的有效性。

现有的研究对图像-文本多模态情感分析存在一定的局限性，如图像特征提取不完整问题、模态之间对齐缺失的问题。

2 网络模型

多视角图文情感分析网络（MITN）的总体架构如图 1 所示。

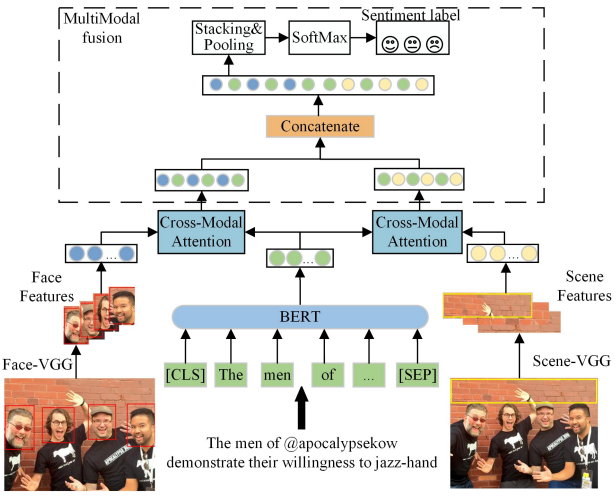


图 1　多视角图文情感分析网络(MITN)总体架构

Fig. 1　Overall architecture of multi perspective image text sentiment analysis network(MITN)

2.1　单模态特征提取

2.1.1　面部表情特征

在图像面部表情特征提取的过程中，给定一个图像 I，在 VGG-19[33]模型上添加了注意力机制得到 VGG-Face，以增强对图像中有脸部表情特征的表达能力，从而获取图像的视觉语义信息。对于每个面部表情，通过平均池化运算得到一个4 096维的特征向量 $\boldsymbol{O}_i$，其中 $i=1,2,\cdots,m$。随后，利用注意力机制来调整特征中的空间权重得到 f_i，通过一个线性投影层将 f_i 转换为一个三维区域特征 F_i，具体计算式如下：

$$f_i=\mathrm{softmax}(ReLU(O_i\cdot W))\cdot O_i \tag{1}$$

$$F_i=W_o f_i+b_o,i\in[1,m] \tag{2}$$

其中，W_o 和 b_o 为可学习参数，F_i 为第 i 个面部表情特征。

2.1.2　场景特征

在图像中提取场景特征图像 S 时，使用 VGG-19 的大型模型框架，并在其中加入空洞卷积(Dilated Convolution)。新改进的模型被称为 Scene-VGG，用于场景特征提取的场景检测器。空洞卷积通过在卷积核中引入膨胀率来增大感受野，从而在不增加参数和计算量的情况下增加输出特征图的尺寸。通过这种方式，可以更好地捕捉到场景中的信息。

在训练阶段，使用 Places 数据集[35]对 Scene-VGG 模型进行训练，将通过迁移学习在后续任务中使用该模型进行场景的情感分析。

将图像 S 输入 Scene-VGG 模型中，并通过空洞卷积和其他卷积层的处理，得到输出特征图。然后，通过平均池化运算，从输出特征图中提取一个 4 096 维的特征向量 $\boldsymbol{g}_i$，其中 $i=1,2,\cdots,p$。接下来，可以通过一个线性投影层将提取的特征向量 $\boldsymbol{g}_i$ 转换为一个三维区域特征 S_i：

$$S_i=W_s g_i+b_s,i\in[1,p] \tag{3}$$

其中，W_s 和 b_s 为可学习参数，S_i 为第 i 个场景特征。

改进后的 Scene-VGG 模型如图 2 所示。

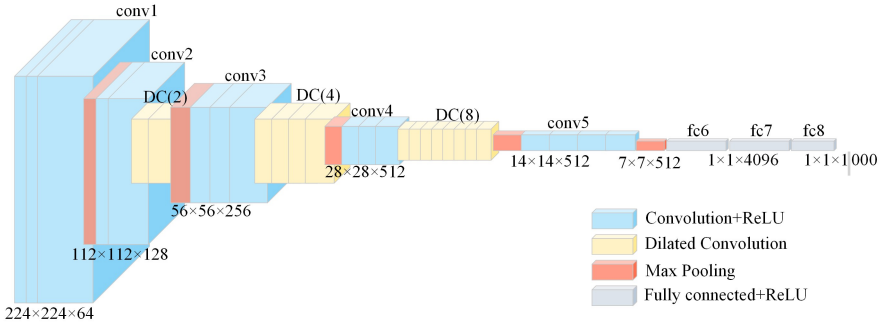


图 2　改进后的 Scene-VGG 模型

Fig. 2　Improved Scene-VGG model

2.1.3　文本特征

本文对于一个有 n 个单词的输入句子 T，使用预先训练过的 BERT-Base[36]将每个单词嵌入到一个 768 维的嵌入向量 $\boldsymbol{x}_i$($i\in[1,n]$)中，使用 Bi-GRU[36]来更好地理解句子中的上下文信息：

$$\overrightarrow{h}_i=\overrightarrow{GRU}(x_i,\overrightarrow{h}_{i-1}),i\in[1,n] \tag{4}$$

$$\overleftarrow{h}_i=\overleftarrow{GRU}(x_i,\overrightarrow{h}_{i+1}),i\in[1,n] \tag{5}$$

其中，$\overrightarrow{h}_i\in\mathrm{Rd}$ 表示前向隐藏状态，$\overleftarrow{h}_i\in\mathrm{Rd}$ 表示后向隐藏状态。最后一个字特征 T_i 被定义为双向隐藏状态的平均值。

$$T_i=\frac{\overrightarrow{h}_i+\overleftarrow{h}_i}{2},i\in[1,n] \tag{6}$$

2.2　跨模态对齐

跨模态对齐指将不同模态(如文本和图像)的信息映射到一个共同的表示空间，以便能够在这个共同的空间中进行有效的比较和分析。由于文本和图像通常表示不同的信息，实现跨模态对齐有助于将它们的特征进行同步，使得模型能够更好地理解它们之间的关联性。在关联的基础上关注了 3 个特征集合：面部表情级特征集合 $O=\{o_1,\cdots,o_m\}$，场景级特征集合 $S=\{s_1,\cdots,s_p\}$ 以及单词级特征集合 $T=\{t_1,\cdots,t_n\}$。

在跨模态对齐机制中，采用了跨模态注意力的方法来实现这种对齐。首先，利用投影矩阵将面部表情、场景和单词的特征映射到一个共同的 k 维空间。然后，计算面部表情-单词注意力矩阵和场景-单词注意力矩阵。

$$A_1=(\hat{W}_o O)(\hat{W}_t T)^{\mathrm{T}} \tag{7}$$

$$A_2=(\hat{W}_s S)(\hat{W}_t T)^{\mathrm{T}} \tag{8}$$

其中，$\hat{W}_o$，$\hat{W}_s$ 和 $\hat{W}_t$ 表示投影矩阵。对于面部表情-单词，场景-单词的注意力矩阵 $\boldsymbol{A}_1\in R^{m\times n}$，$\boldsymbol{A}_2\in R^{p\times n}$。这些注意力矩阵通过衡量每个面部表情对应的单词以及每个场景对应的单词之间的关联程度，实现了不同模态之间的信息对齐。图 3 为面部表情与场景的跨模态对齐模块。

为了推断局部片段之间的潜在对齐，对注意力矩阵进行了归一化处理。这一步骤通过将注意力集中在每个区域与单

词之间的关系上，加强了局部的语义对齐。具体而言，计算了面部表情-单词归一化注意力矩阵和场景-单词归一化注意力矩阵。

$$\bar{A}_1=\text{softmax}\left(\frac{A_1}{\sqrt{k}}\right) \tag{9}$$

$$\bar{A}_2=\text{softmax}\left(\frac{A_2}{\sqrt{k}}\right) \tag{10}$$

其次，利用这些归一化矩阵来聚合每个区域的单词特征。面部表情级特征聚合（见式（11））以及场景级特征聚合（见式(12)）能够从不同角度探索图像与文本之间的相互作用，进而捕捉到它们之间的语义关系。

$$U_1=\bar{A}_1\cdot W \tag{11}$$

$$U_2=\bar{A}_2\cdot W \tag{12}$$

其中，U_1，U_2的第 i 行表示对应于第 i 个区域的交互文本特征。因此，U 可以通过对齐区域和单词来探索图像和句子之间的相互作用。

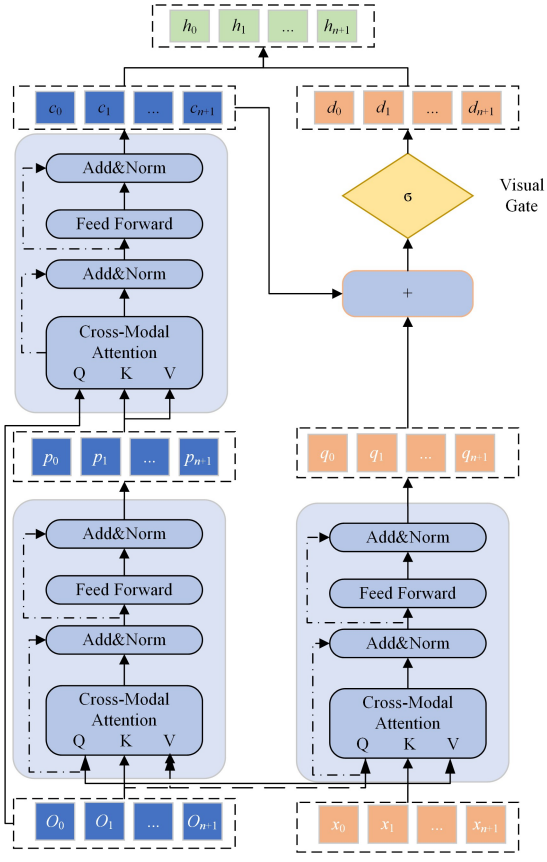


图 3 面部表情与场景跨模态对齐模块

Fig. 3 Facial expression and scene cross-modal alignment module

2.3 多模态特征融合

本文采用混合融合的策略，通过堆叠池模块来融合多模态特征。图 4 清楚地展示了这一过程。

该方法的核心在于在特征融合过程中引入混合融合，以弥补早期融合[37]和晚期融合[38]的不足。将面部表情文本交互特征和场景文本交互特征U_1，U_2进行连接，获得一个融合后的特征Q^i表示，表达式如下：

$$Q^i=f_{\text{concat}}(U_1^i,U_2^i) \tag{13}$$

随后，该特征Q^i被输入到一个多层（四层）感知器网络中。为了更好地学习不同模态特征之间的交互作用，堆叠了多个完全连接的层，其中包括FC_2，FC_3和FC_4。为了完成深度融合，使用了卷积池操作。因此，堆叠池模块的结构可以表示为一个 512×3 的矩阵，通过多个 1×1×3 的卷积核进行卷积操作，之后进行合并，从而得到深度融合特征。卷积运算和最大池化运算的过程如下：

$$F^i=f_{\text{pooling}}(f_{\text{conv}}(F_2^i,F_3^i,F_4^i)\theta^{\text{conv}}) \tag{14}$$

最终，融合后的特征被输入到一个 softmax 分类器中，用于生成最终的标签。具体的表述为：

$$L^i=f_{\text{Softmax}}(F^i;\theta^{\text{softmax}}),L^i\in R^C \tag{15}$$

其中，θ^{conv}和θ^{softmax}是卷积层和软层的参数。本文采用交叉熵损失函数，因此模型损失函数可以表示为：

$$J(\theta^{\text{conv}},\theta^{\text{softmax}})=\sum_{i=0}^{u-1}-\log(L^i)=f_{\text{softmax}}(f_{\text{pooling}}(f_{\text{conv}}(f_{\text{multi-layer}}(U_1|U_2));\theta^{\text{conv}});\theta^{\text{softmax}}) \tag{16}$$

其中，$f_{\text{multi-layer}}$是一个多层感知器操作，“|”表示连接操作。

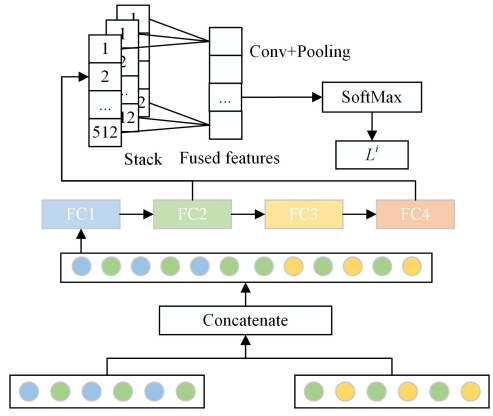


图 4 使用堆叠池模块融合不同特性的过程

Fig. 4 Process of fusing different features with stacked pool modules

混合融合的优点在于其灵活性，它能够同时在特征和决策层面融合信息。相对于早期融合，它能够更好地保留语义信息，同时又避免了晚期融合可能出现的信息丢失问题。此外，通过在多个层次上进行特征融合，它能够更好地捕捉模态之间的交互，提升多模态情感分类的泛化能力。

3 实验

3.1 数据集

为了全面评估所提模型的性能，本文在两个公开的多模态情感分析数据库上进行了实验评估，分别是 MVSA-Single 和 MVSA-Multiple[39]，其样本是从 Twitter 上收集的图文评论。MVSA-Single 中的每个样本由 1 位标注者进行标注只包含 1 个情感标注标签，共 4869 个图像-文本对；MVSA-Multiple 中的每个样本由 3 位标注者进行标注，包含 3 个情感标注标签，共 19598 个图像-文本对。在进行实验前，对这两个数据集进行了预处理。如果一幅图像的标签是积极的（或消极的），而其对应的文本标签是中性的，将此对图像和文本标签的情感极性判定为积极（或消极）。同时，为了确保标签的一致性，排除了在图像和文本标签上不一致的样本。在排除之后两种数据集的总数如表 1 所列。

表 1 列出了两个 MVSA 数据集的统计数据。然而，值得注意的是，这些数据集中不同类别的分布存在显著的不平衡情况。具体而言，MVSA-Single 和 MVSA-Multiple 数据集的数据分布存在差异。若不进行数据抽样处理，使用包含较小类别数据的数据集将会在训练过程中面临困难，甚至可能导

致所有数据被错误地归类为同一类别，从而导致分类器失效。

表 1　在两个数据集上的不同情感标签的数量

Table 1　Number of different emotional labels on two datasets

Dataset	Positive	Neutral	Negative	Total
MSVA-Single	2682	469	1357	4508
MSVA-Multiple	11316	4406	1296	17018

因此，为了保证实验的可行性和结果的可靠性，采用了随机上采样的方法来处理每个 MVSA 数据集中最小类别的样本。通过这一方法，能够减轻数据不平衡对实验结果的影响，使得训练过程更加平衡且有效。

Places 数据集，其中有 1000 万张场景图片，434 个场景语义类别，包括了人类可以在地球上经历的 98%的场景。为了进行场景特征提取，使用 Scene-VGG 模型，并通过对该数据集进行预训练来获得更有表现力的场景表示。

3.2　对比实验

为了充分验证模型的性能，本文与以下图片文本模态情绪分析基线模型进行了比较。

ImaText-IST[40]：通过图像翻译模块，将图像转化为情感丰富的图像描述，包括积极、中性和消极情感。这些图像描述与数据集中的文本进行情感相关性分析，提升图像情感理解的准确性。同时，结合图像语义描述、目标及文本进行情感预测，采用特征融合和辅助语句的方式进行情感分析。

MBAH[41]：基于深度模型提取的图像和文本特征，通过双向注意力机制引入跨模态信息，学习模态间关联。底层特征与语义特征经注意力计算融合，形成共享表征，经多层感知器分类。MBAH 模型采用后期融合技术，结合自注意力模型搜寻决策权值，实现最终分类决策。

MVAN[42]：在多视角注意网络的基础上开发了一个多模态情绪分析模型，该模型利用记忆网络模块迭代获取语义图像-文本特征。与其他图像-文本多模态情绪分析研究相比，该方法取得了最先进的性能。

ITIN[43]：引入了一个跨模态对齐模块来捕获区域-字的对应关系，在此基础上，多模态特征通过一个自适应的跨模态门控模块进行融合。此外，考虑到上下文信息在情绪分析中的互补作用，整合了个体-模态的上下文特征表示，以实现更可靠的预测。

CLMLF[44]：研究编码文本和图像以获得隐藏表示，用多层融合模块对齐和融合二者标记级特征。并设计标签和数据对比学习任务，帮助模型学习多模态情感相关特征。

表 2　不同方法在 MVSA 数据集上的 Acc 和 F1 分数对比

Table 2　Comparison of Acc and F1 scores of different methods on MVSA dataset

Method	MVSA-Single		MVSA-Multiple	
	Accuracy	*F*1	*Accuracy*	*F*1
ImaText-IST[40]	0.7223	0.7230	0.7153	0.7133
MBAH[41]	0.7355	0.7343	0.7256	0.7253
MVAN[42]	0.7298	0.7298	0.7236	0.7230
ITIN[43]	0.7519	0.7497	0.7352	0.7349
MITN(Ours)	0.7624	0.7612	0.7313	0.7303
CLMLF[44]	0.7633	0.7646	0.7350	0.7383

CLMLF 模型在多模态情感分析任务中表现出了较高的准确率，主要归因于其对比学习模块在多模态特征融合方面的优越。然而，CLMLF 的训练需要大量数据并且泛化能力差，对于一些大样本数据，相同特征会归为一类的思想使得其准确率高，但对于小样本数据，不同的特征无法归为一类，因此准确率不高。本研究的 MITN 模型同时注重大样本和小样本数据，通过图像特征的多视角，从面部表情和场景两个角度出发分别和文本信息交互，通过引入注意力机制、空洞卷积、迁移学习等方法，提高了模型的泛化能力，更全面地捕捉多模态任务的关键特征。

3.3　实验细节

在此次实验中，数据集按 8∶1∶1 的分割比例随机分为训练集、验证集和测试集。所提出的 MITN 由 Adam 优化器进行更新。为避免过拟合，在全连接层之后使用 Dropout（值设为 0.1），设置学习率为 0.001，epoch 为 100。考虑到两个数据集中的样本数量不同，本文将 MVSA-Single 的批大小设置为 64，MVSA-Multiple 的批大小设置为 128。

由于 MVSA 数据集中的数据大多是不平衡的，因此绘制了 MITN 模型在 MVSA-Single 和 MVSA-Multiple 上的 P-R 曲线图（精准度-召回率曲线），如图 5 所示。这样有助于了解模型的实际效果和作用，也能够以此改进模型性能。其中，使用精准度（precision）和召回率（recall）作为评价模型性能指标的数据，其定义如下：

$$Precision=\frac{TP}{TP+FP} \tag{17}$$

$$Recall=\frac{TP}{TP+FN} \tag{18}$$

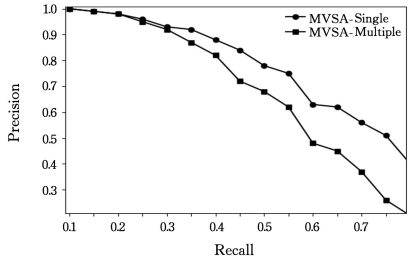


图 5　模型在 MVSA-Single 和 MVSA-Multiple 上的 *P-R* 曲线

Fig. 5　*P-R* curves of MIyTN model on MVSA-Single and MVSA-Multiple

从曲线结果可以看出，MITN 模型对学习图文双模态情感分类任务展现了有效结果。

以 MVSA-Multiple 为研究对象，为了更好地看出模型训练效果的优劣，绘制了模型训练和验证过程的损失收敛曲线。图 6 给出了损失值随 epoch 次数的变化过程，它随着数据量的增大而减小，训练过程的损失大约在 52 次后趋向平稳，损失值在 0.02 左右，损失值越小表示模型收敛程度越高。

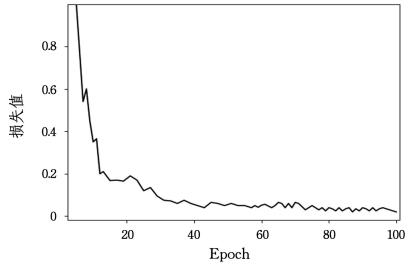


图 6　模型训练和验证过程在 MVSA-Multiple 上的损失收敛曲线

Fig. 6　Loss convergence curve during model training and validation process in MVSA-Multiple

3.4　消融实验

为评估 MITN 中模块的减少对多模态差异的影响，本研

究在 MVSA 数据集下进行了消融研究，以探索最佳组合。本文进行了 6 组消融实验，分别排除了模型中的场景特征（w/o Scene）、面部表情特征（w/o Face）、跨模态注意力模块（w/o CMA）和混合融合模块（w/o MFF），以及仅面部表情特征（only Face）和仅场景特征（only Scene）。

通过这些实验，评估了每个组成部分对整体模型性能的贡献，并深入了解了它们在图文多模态情感分析任务中的作用。表 3 列出了这些消融实验结果。

表 3 在 MVSA 数据集上的消融对比实验

Table 3 Comparison of ablation experiments on MVSA dataset

Datasets	Model	Accuracy	F1
MVSA-Single	MITN w/o Scene	0.7142	0.7120
	MITN w/o Face	0.6875	0.6802
	MITN only Face	0.6998	0.6995
	MITN only Scene	0.6686	0.6680
	MITN w/o CMA	0.7263	0.7199
	MITN w/o MFF	0.7426	0.7425
	MITN	0.7624	0.7612
MVSA-Multiple	MITN w/o Scene	0.7055	0.7045
	MITN w/o Face	0.6996	0.6966
	MITN only Face	0.6863	0.6821
	MITN only Scene	0.6598	0.6548
	MITN w/o CMA	0.7130	0.7168
	MITN w/o MFF	0.7201	0.7201
	MITN	0.7313	0.7303

在 MITN only Scene 实验中，发现当模型中不包含面部表情和文本信息时，仍然能够检测出情感。在实际应用中很难保证图像和文本同时存在的情况或数据集中的文本无法识别的问题，此时其他研究者会在预处理阶段在数据集中删除此类情况，从而避免影响结果。因此，这表明现有模型在这方面仍有不足。同时，这也验证了 MITN 具有较好的泛化性能。

在模型中同时包含图像中的面部表情特征和场景特征时，其性能表现达到最佳水平。通过跨模态注意力模块和混合融合模块，文本和图像数据得以相互关联，从而增强了情感分析过程中的学习效果。从实验结果可以看出，完整的 MITN 模型在性能上达到了最佳水平。

3.5 样本分析

表 4 中，通过面部表情检测器观察到人物微笑，场景为自然景观呈积极情感。结合文本中的“Enjoying”，可以明确地推断出该图像传达了积极情感。这一结论不仅基于图像中的视觉信息，也得到了文本信息的支持。

表 4 预测样例 1

Table 4 Prediction example 1

Enjoying a stroll #butchartgardens #kilts

MVSA-Single 标注情感类别：积极

MITN 预测情感类别：面部表情（积极）+场景（中性）+文字（积极）=积极

ITIN 预测情感类别：面部表情（积极）+文字（积极）=积极

CLMLF 预测情感类别：面部表情（积极）+文字（积极）=积极

表 5 中，一辆空火车没有人物，面部表情检测器无法直接检测到面部表情的情感。然而，通过对图片的场景情感分析呈现了一个灰褐色调的空车厢，得出其为消极情感。同时，文本中的 empty 也表达了消极情感。因此，最终的情感判断为消极。这一案例显示了 MITN 模型在处理缺少直接面部表情情感的场景时，仍然能够准确捕捉到情感信息。

表 5 预测样例 2

Table 5 Prediction example 2

See youFrance? # train # empty # Eurostar # London # Travel # trip # luggage # intern

MVSA-Single 标注情感类别：消极

MITN 预测情感类别：面部表情（中性）+场景（消极）+文字（消极）=消极

ITIN 预测情感类别：面部表情（中性）+文字（中性）=中性

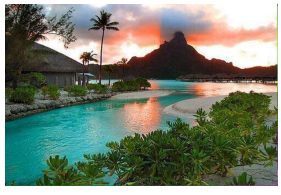
CLMLF 预测情感类别：面部表情（中性）+文字（中性）=消极

通过上述实例分析，证明了 MITN 模型在不同样本下都能够做出较为准确的情感预测。这归功于模型对文本模态和视觉模态信息的有效捕捉和整合。这种能力进一步证明了 MITN 模型的多样性和适用性，使其在实际应用中具有显著的优势。

然而，为了更全面地评估 MITN 模型的性能，还需要考虑一些特殊情况。表 6 列出了一张风景照，其中未出现明显的人脸表情和缺少直接面部表情，并且文本也没有体现情感。因此，在 ITIN 及 CLMLF 上，判断的情感极性都为中性。然而，MITN 模型在对图片进行场景特征提取后，判别为积极情感，这与最后数据集给出的情感类别一致。这一结果强调了 MITN 模型的泛化能力，使其在面对不同数据分布时仍能保持稳定的性能。

表 6 预测样例 3

Table 6 Prediction example 3



RT @TheBucktList: Bora Bora??

MVSA-Single 标注情感类别：积极

MITN 预测情感类别：面部表情（中性）+场景（积极）+文字（中性）=积极

ITIN 预测情感类别：面部表情（中性）+文字（中性）=中性

CLMLF 预测情感类别：面部表情（中性）+文字（中性）=中性

结束语 本文提出了一种基于图像的两种视图（面部表情视图和场景视图）与文本信息相结合的图像文本网络（MITN）模型，用于多模态情绪分析。通过迁移学习，有效捕捉场景中的情感极性。跨模态注意力机制模块实现了图像中面部表情和场景区域与文本词的对齐，从而实现细粒度的跨模态情感分析，且提高了模型的泛化能力。堆叠池模块构建了多模态特征融合模块，这两者提升了多模态情感分析的质量。在公开数据集 MVSA-Single 和 MVSA-Multiple 上的实验结果验证了 MITN 模型优于大部分基线模型。随着社交

媒体中短视频的活跃，在未来的工作中，考虑将把音频加入多模态情感分析中，面向这些数据的情感分析研究仍需进一步探索。

参 考 文 献

[1] GIATSOGLOU M，VOZALIS M G，DIAMANTARAS K，et al. Sentiment analysis leveraging emotions and word embeddings [J]. Expert Systems with Applications，2017，69：214-224.

[2] SINGH V，RAM M，PANT B. Identification of zonal-wise passenger's issues in Indian railways using latent Dirichlet allocation（LDA）：A sentiment analysis approach on tweets[M]// Mathematics Applied in Information Systems. 2018.

[3] CHATURVEDI I，RAGUSA E，GASTALDO P，et al. Bayesian network based extreme learning machine for subjectivity detection[J]. Journal of The Franklin Institute，2018，355(4)：1780-1797.

[4] BANDHAKAVI A，WIRATUNGA N，MASSIES，et al. Lexicon generation for emotion detection from text[J]. IEEE intelligent systems，2017，32(1)：102-108.

[5] PORIA S，CAMBRIA E，BAJPAIR，et al. A review of affective computing：From unimodal analysis to multimodal fusion[J]. Information Fusion，2017，37：98-125.

[6] HUANG Y，DU C，XUE Z，et al. What Makes Multimodal Learning Better than Single（Provably）[J]. Advances in Neural Information Processing Systems，2021，34：10944-10956.

[7] DENG D，ZHOU Y，PI J，et al. Multimodal utterance-level affect analysis using visual，audio and text features[J]. arXiv：1805.00625，2018.

[8] SHUTOVA E，KIELA D，MAILLARD J. Black holes and white rabbits：Metaphor identification with visual features[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2016：160-170.

[9] YU Y，LIN H，MENG J，et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks [J]. Algorithms，2016，9(2)：41.

[10] LIU H Y，HU Z G，PENG D L. The interaction of emotion and language processing [J]. Advances in Psychological Science，2009，17(4)：714.

[11] ORTIS A，FARINELLA G M，BATTIATO S. An Overview on Image Sentiment Analysis：Methods，Datasets and Current Challenges[J]. ICETE(1)，2019：296-306.

[12] COLOMBO C，DEL BIMBO A，PALA P. Semantics in visual information retrieval[J]. IEEE Multimedia，1999，6(3)：38-53.

[13] SCHMIDT S，STOCK W G. Collective indexing of emotions in images. A study in emotional information retrieval[J]. Journal of the American Society for Information Science and Technology，2009，60(5)：863-876.

[14] BORTH D，JI R，CHEN T，et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]// Proceedings of the 21st ACM International Conference on Multimedia. 2013：223-232.

[15] YOU Q，JIN H，LUO J. Visual sentiment analysis by attending on local image regions[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[16] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv：1409. 1556，2014.

[17] ZHOU B，LAPEDRIZA A，KHOSLA A，et al. Places：A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(6)：1452-1464.

[18] ASGHAR M Z，KUNDI F M，AHMAD S，et al. T-SAF：Twitter sentiment analysis framework using a hybrid classification scheme[J]. Expert Systems，2018，35(1)：e12233.

[19] HAMOUDA A，ROHAIM M. Reviews classification using sentiwordnet lexicon[C]// World Congress on Computer Science and Information Technology. sn，2011，23：104-105.

[20] TANG D，WEI F，QIN B，et al. Coooolll：A deep learning system for twitter sentiment classification[C]// Proceedings of the 8th International Workshop on Semantic Evaluation（SemEval 2014）. 2014：208-212.

[21] YANG Z，YANG D，DYER C，et al. Hierarchical attention networks for document classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2016：1480-1489.

[22] CHIONG R，FAN Z，HU Z，et al. A sentiment analysis-based machine learning approach for financial market prediction via news disclosures[C]// Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2018：278-279.

[23] XU J，HUANG F，ZHANG X，et al. Visual-textual sentiment classification with bi-directional multi-level attention networks [J]. Knowledge-Based Systems，2019，178：61-73.

[24] HUANG F，ZHANG X，ZHAO Z，et al. Image-text sentiment analysis via deep multimodal attentive fusion[J]. Knowledge-Based Systems，2019，167：26-37.

[25] XU J，LI Z，HUANG F，et al. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations[J]. IEEE Transactions on Industrial Informatics，2020，17(4)：2974-2982.

[26] YANG J，YU Y，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（Volume 1：Long Papers）. 2023：7617-7630.

[27] FAN F，FENG Y，ZHAO D. Multi-grained attention network for aspect-level sentiment classification[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018：3433-3442.

[28] ZHANG L，ZHANG X，PAN J. Hierarchical cross-modality semantic correlation learning model for multimodal summarization [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022，36(10)：11676-11684.

[29] PORIA S，CAMBRIA E，HAZARIKA D，et al. Multi-level multiple attentions for contextual multimodal sentiment analysis [C]// 2017 IEEE International Conference on Data Mining（ICDM）. IEEE，2017：1033-1038.

[30] ZADEH A，CHEN M，PORIA S，et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv：1707. 07250，2017.

[31] AREVALO J, SOLORIO T, MONTES-Y-GÓMEZ M, et al. Gated multimodal units for information fusion[J]. arXiv: 1702.01992, 2017.

[32] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors [J]. arXiv: 1806.00064, 2018.

[33] YOU Q, LUO J, JIN H, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia[C]// Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. 2016: 13-22.

[34] SIMONYAN K, ZISSERMANA. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[35] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 Million Image Database for Scene Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 40(6): 1452-1464.

[36] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// NAACL. 2019: 4171-4186.

[37] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// Proceedings of the International Conference on Learning Representations. 2015.

[38] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[39] NIU T, ZHU S, PANG L, et al. Sentiment analysis on multiview social data, [C]// Proceedings of the International Conferenceon Multimedia Modeling. 2016: 15-27.

[40] HUANG J, WANG Y. Emotional Analysis Method for Image Text Fusion Based on Image Semantic Translation [J]. Computer Engineering and Applications, 2023, 59(11): 180-187.

[41] HUANG H Z, MENG Z Q. Multimodal sentiment classification method based on bidirectional attention mechanism [J]. Computer Engineering and Applications, 2021, 57(11): 9.

[42] YANG X, FENG S, WANG D, et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE Transactions on Multimedia, 2020, 23: 4014-4026.

[43] ZHU T, LI L, YANG J, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE Transactions on Multimedia, 2022, 25: 3375-3385.

[44] LI Z, XU B, ZHU C, et al. CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection [J]. arXiv: 2204.05515, 2022.

GAO Weijun, born in 1973, associate professor. His main research interests include software engineering, natural language processing, multimodal sentiment analysis.

SUN Zibo, born in 1998, graduate student. His main research interest is multimodal sentiment analysis multimodal sentiment analysis