浙江理工大学学报，2022，47(1)：85-94
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.01.012

基于多层跨模态注意力融合的图文情感分析

陈巧红，孙佳锦，孙　麒，贾宇波
（浙江理工大学信息学院，杭州　310018）

摘　要：针对现有图文情感分析模型仅考虑图像高层特征与文本特征的联系，而忽视图像低层特征的问题，提出了一种基于多层跨模态注意力融合(Multi-level cross-modal attention fusion，MCAF)的图文情感分析模型。该模型首先将 VGG13 网络外接多层卷积，以获取不同层次的图像特征，并使用 BERT 词嵌入与双向门控循环网络(Gated recurrent unit，GRU)网络获取文本情感特征；然后将提取后的多层图像特征与文本特征进行注意力融合，得到多组单层文本-图像注意力融合特征，并将其通过注意力网络分配权重；最后将得到的多层文本-图像注意力融合特征输入全连接层，得到分类结果。在公开的 MVSA 和 Memotion-7k 数据集上进行实验，结果显示：与图文情感分析基线模型相比，基于多层跨模态注意力融合的图文情感分析模型的准确率和 F1 值在 MVSA 数据集上分别提升 2.61%和 3.56%，在 Memotion-7k 数据集上分别提升 3.25%和 3.63%。这表明该模型能够有效提高图文情感分类性能。

关键词：图文情感分析；门控循环网络；注意力机制；跨模态融合；多层图像特征抽取
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 01-0085-10

Image-text sentiment analysis based on multi-layer cross-modal attention fusion

CHEN Qiaohong, *SUN Jiajin*, *SUN Qi*, *JIA Yubo*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In view that the existing image-text sentiment analysis models only consider the relationship between high-level features of image-text features with no consideration of the low-level features of image, an image-text sentiment analysis model based on multi-layer cross-modal attention fusion (MCAF) was proposed. The model was firstly connected the VGG13 network with a multi-layer convolution to obtain different levels of image features, and then the text sentiment features were obtained using BERT word embedding and bidirectional GRU network. Subsequently, the attention fusion was performed on the extracted multi-layer image features and text features. Then, multiple sets of single-layer text-image attention fusion features were obtained, which were used for weights allocation through the attention network. Finally, the obtained multi-layer text-image attention fusion features were inputted into the full-connection layer to get classification results. After the experiment on the public MVSA and Memotion-7k datasets, the results revealed that compared with the image-text sentiment baseline analysis model, the accuracy and F1 value of the image-text sentiment analysis model based on multi-layer cross-modal attention fusion increased by 2.61% and 3.56% on the MVSA datasets, respectively, and also increased by 3.25% and 3.63% on the Memosion-7k datasets, indicating that this model could effectively improve

收稿日期：2021－05－23　　网络出版日期：2021－10－11
基金项目：浙江理工大学中青年骨干人才培养经费项目
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习方面的研究。

the performance of image-text sentiment classification.

Key words: image-text sentiment analysis; gated recurrent unit (GRU) network; attention mechanism; cross modal fusion; multi-layer image feature extraction

0 引 言

随着社交网络的快速发展,人们渐渐习惯用多种模态的组合数据来表达自己的情感。对互联网上存在的海量图文信息进行情感分析,有利于了解人们对某些事件的态度和看法,协助政府机构更高效地监管网络内容、了解民情民意,进行正确地舆论引导;有利于帮助企业了解顾客的喜好与需求,从而改进产品,增加收益。目前,图文情感分析依然是一项复杂的任务。首先,图像与文本模态所蕴含的情感信息并不总是一致的,对图文数据进行情感分析需要准确地提取不同模态的情感特征;其次,用于表达不同模态数据的低层特征维度往往不同,与传统的单模态情感分析模型相比,多模态情感分析要求准确匹配文本模态与图像模态特征,以有效挖掘图像与文本之间的情感共现[1]。

目前多模态数据融合的方式主要有三种,即前期融合[2-4]、后期融合[5-6]以及混合融合[7-9]。范涛等[3]提出一种基于前期融合的模型,将经过卷积的图像特征与双向 LSTM 网络处理后的文本特征进行拼接,以学习两种模态低级特征之间的相关性。缪裕青等[6]将不同模态数据分别输入训练好的单模态情感分类器,输出打分(决策),再进行加权融合,这种基于后期融合的模型,通过加权不同模态独立的情感得分,使模型选取置信度更高的模态结果作为多模态数据整体的预测情感,能取得优于单模态模型的性能,但是因为不同模态间的数据缺少交互,难以学习每种模态之间的相关性。谢豪等[9]采用混合融合的方法,先将文本特征与图像特征相互进行引导注意力,得到两个融合模态的情感分数,再使用加权策略对两个模型的情感分类得分进行后期融合,充分挖掘了图文之间的双向多层次语义关联。Zadeh 等[10]采用张量融合(Tensor fusion network, TFN)的方式,将不同模态的特征通过笛卡儿积获取高维张量,该方法的性能比传统的基于模态拼接或点乘的方法有明显提升,缺点是计算复杂度较高。郭可心等[11]提出了一种空间注意力网络,利用文本特征对卷积神经网络中的图像特征构建空间注意力,以挖掘文本特征与图像特征的情感共现。

在文本情感分析方面,Li 等[12]使用 Google 发布的 BERT 预训练模型,从股票投资者发布的在线评论中提取情绪值,再将这些情绪值通过注意力加权,来计算投资者的情绪指标,因 BERT 模型具有优秀的语言理解能力,该方法的准确率要高于 LSTM 和 SVM 模型。Chen 等[13]提出了一种非序列文本情感分析模型,将经过 Glove 词嵌入后的文本向量进行二维卷积,有效提取了文本情感,验证了卷积神经网络在自然语言处理方面的可行性。

在图像情感分析方面,Campos 等[14]通过使用卷积网络对图像进行特征提取,将提取的高层语义特征通过前馈层进行情感分类,取得了高于传统机器学习方法的准确率,但图像的情感并不只与高层语义相关。研究表明,人们对图像情感的认知,会受图像不同层次特征的影响[15],当人们看到图像中有笑脸时,通常会产生积极的情感,当看到色彩亮丽的图像时,其产生的情感往往也是积极的。然而,传统的卷积神经网络模型主要是针对目标分类而设计的,通过不断叠加卷积层来提取图像的抽象视觉特征[16],这类模型往往被用以识别物体,但并不能有效地提取色彩、纹理等低层图像特征。基于该问题,Rao 等[17]提出了多层深度表示网络,将卷积网络 AlexNet 的不同卷积层进行外接,以提取各层的图像特征,再将这些特征进行拼接,有效地利用了不同层次的图像情感信息。

上述对图文情感分析的研究,通过对特征提取与模态融合的设计,在某些场景下可以有效地联系并理解图文信息,但仍然缺少对多模态情感共现及其内部原理的研究,其模型大多将卷积神经网络作为提取图像高层语义的黑盒[11],并未有效挖掘图像的低层特征与文本特征之间的情感联系。例如,当所识别的图文样例中,图像为面带微笑的人脸时,上述模型往往能准确地识别出该条图文样例的情感;但当图像为风景照或者抽象画时,图像的情感不再由某一实体来表现,而是通过图像的整体色彩、画风来展示情感,对于此类图文样例,上述模型很难有效提取图像的情感色彩,降低了对整体情感的识别率。

为了提升情感分析模型对不同类型图文数据的泛化性能,探究通过抽取多层图像特征与注意力融合以提升模型对不同类型图文样例的情感分析准确

率的可行性，本文提出了一种基于多层跨模态注意力融合的图文情感分析模型。该模型使用双向门控循环网络(Gated recurrent utils，GRU)实现对文本情感特征的表示，能够增强文本特征内部的联系，使其具有上下文联系性；同时利用外接多层卷积的 VGG13 网络，来获取不同层次的图像特征；通过将文本特征分别与多层图像特征进行注意力计算，获取多层图文融合特征，并将多层图文融合特征输入多层感知机与 Softmax 层进行情感分类。由于该模型有效地融合了多层图像特征与文本特征，其情感分类效果将得到提升。

1　模型构建

1.1　整体流程

本文提出的多层跨模态注意力融合模型采用端到端的训练方式，主要包含以下四个模块：文本特征提取模块、图像特征提取模块、跨模态特征融合模块和情感分类输出模块。首先对输入的英文句子进行分词，将每个单词通过 BERT 词嵌入获取对应的词向量，并将词向量输入双向 GRU 网络获取句子的情感特征。在多层图像特征提取模块中，将图像输入外接多层卷积的 VGG13 网络，获得不同层次的图像情感特征。在跨模态特征融合模块中，首先将句子情感特征与各层图像情感特征进行注意力融合，获取多个单层图像－文本注意力关联特征；其次通过加性注意力，将注意力矩阵与各层图像－文本注意力关联特征进行计算，获得不同层特征的注意力权重；然后将各层图像－文本注意力关联特征经过加权求和，得到多层图像－文本注意力关联特征；最后将其输入多层感知机与 Softmax 层得到情感分类结果。本文设计的多层跨模态注意力融合模型的整体流程如图 1 所示。

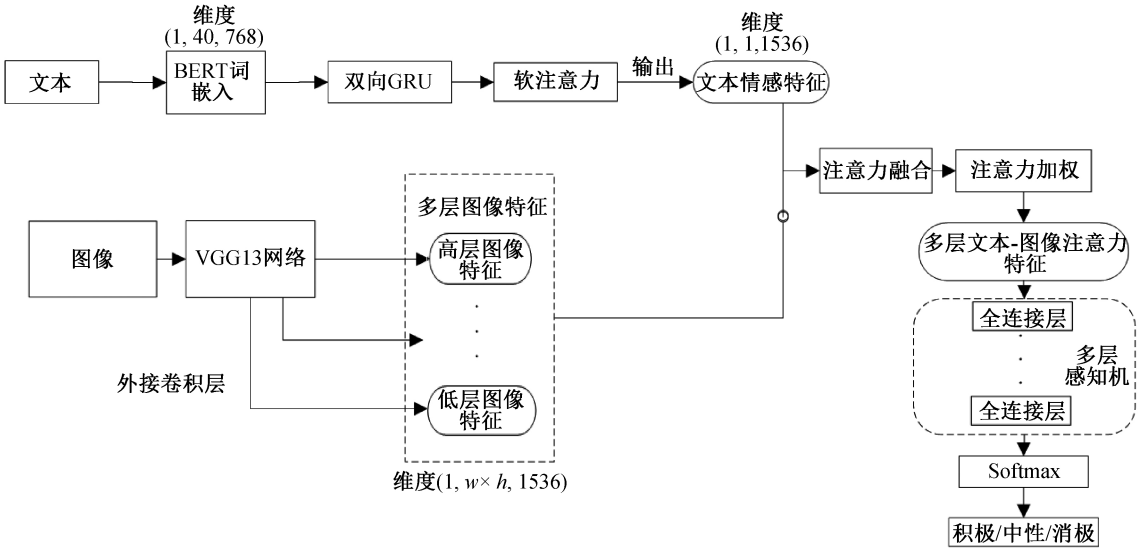

图 1　多层跨模态注意力融合模型

1.2　文本特征提取

给定一个图文对$(\boldsymbol{P}_i,\boldsymbol{T}_i)$，对于文本模态 $\boldsymbol{T}_i$，设第 i 条文本由个 n 单词构成，则该条文本可表示为 $\boldsymbol{T}_i=(\boldsymbol{t}_{i,1},\cdots,\boldsymbol{t}_{i,k},\cdots,\boldsymbol{t}_{i,n})$，其中 $\boldsymbol{t}_{i,k}$ 表示该条文本的第 k 个单词。在文本特征提取阶段，有两种方式来对文本特征进行表达，分别是独热编码(One-hot)与词向量表示。在独热编码表示中，每个词将会用一个长向量表示，容易导致向量稀疏并使得模型难以训练。词嵌入技术能够将单词映射至低维特征空间，成为稠密词向量，经过词嵌入之后的单词，其特征的余弦相似度越小，单词之间的相关性就越大。传统的词嵌入技术，如 Glove 和 Word2vec，用一个固定的向量来表示一个具体的单词，预训练完词向量之后可以直接通过查找词向量映射表来使用，其缺点是固定的词向量无法识别具有多个含义的单词。为了实现不同语境中一个单词可以有不同的意思，本文使用 BERT[18] 词嵌入，根据文本信息获取具体的词向量：$\boldsymbol{T}_i^{\text{word}}=\text{BERT}(\boldsymbol{T}_i)$，将每个单词嵌入成一个 d 维词向量 $\boldsymbol{t}_{i,k}^{\text{word}}\in\mathbf{R}^{\text{d}}$，即 $\boldsymbol{T}_i^{\text{word}}=(\boldsymbol{t}_{i,1}^{\text{word}},\cdots\boldsymbol{t}_{i,k}^{\text{word}},\cdots\boldsymbol{t}_{i,n}^{\text{word}})$。对于每一个词嵌入向量，模型使用了 GRU[19]，进行进一步编码，与传统循环神经网络(Recurrent Neural Networks，RNNs)相比，GRU 通过添加更新门与重置门，能够对输入的时序信息选择性的保留与遗忘，从而捕捉到重要的长期信息。该模块结构如图 2 所示。假定 k 时刻 GRU 接受单词嵌入 $\boldsymbol{t}_{j,k}^{\text{word}}$作为输入，上一时刻输出为 $\boldsymbol{h}_{i,k-1}$，并输

出一个新的隐藏状态向量 $\boldsymbol{h}_{i,k}$，计算过程如式(1)所示：

$$\begin{cases} z_{i,k}=\sigma(\boldsymbol{W}_{\text{update}}\boldsymbol{t}_{i,k}^{\text{word}}+\boldsymbol{U}_{\text{update}}\boldsymbol{h}_{i,k-1}), \\ r_{i,k}=\sigma(\boldsymbol{W}_{\text{reset}}\boldsymbol{t}_{i,k}^{\text{word}}+\boldsymbol{U}_{\text{reset}}\boldsymbol{h}_{i,k-1}), \\ \tilde{\boldsymbol{h}}_{i,k}=\tanh(\boldsymbol{W}\boldsymbol{t}_{i,k}^{\text{word}}+\boldsymbol{U}_{\text{text}}(r_{i,k}\times\boldsymbol{h}_{i,k-1})), \\ \boldsymbol{h}_{i,k}=(1-z_{i,k})\times\boldsymbol{h}_{i,k-1}+z_{i,k}\times\tilde{\boldsymbol{h}}_{i,k}, \\ \boldsymbol{h}_{i,k}=\text{GRU}(\boldsymbol{t}_{i,k}^{\text{word}},\boldsymbol{h}_{i,k-1}) \end{cases} \tag{1}$$

其中：$\boldsymbol{W}_{\text{update}}$、$\boldsymbol{U}_{\text{update}}$、$\boldsymbol{W}_{\text{reset}}$、$\boldsymbol{U}_{\text{reset}}$、$\boldsymbol{W}$、$\boldsymbol{U}_{\text{text}}$都是对应特征的映射矩阵，其参数由模型训练生成，$z_{i,k}$为更新门，表示以前的状态是否需要更新；$r_{i,k}$为重置门，类似于 LSTM 的忘记门，代表以前的状态是否需要重置，$\tilde{\boldsymbol{h}}_{i,k}$表示待选隐藏状态向量，其值由 $r_{i,k}$决定，判断是否放弃之前的隐藏状态 $\boldsymbol{h}_{i,k-1}$，最后由更新门 $z_{i,k}$决定所产生的新隐藏状态向量 $\boldsymbol{h}_{i,k}$。由于某一单词的情感受前后文内容影响，该模型引入双向 GRU 机制(Bi-directional GRU)，将前向 GRU 和后向 GRU 生成的隐藏状态向量进行拼接，得到最终的单词特征表示：

$$\boldsymbol{h}_{i,k}=\overrightarrow{\boldsymbol{h}}_{i,k}\oplus\overleftarrow{\boldsymbol{h}}_{i,k},$$

一个句子里的每个单词所蕴含的情感信息和强度是不对等的，某些重要的单词应该在文本情感中占更大的比重，为了重新分配不同单词所占的情感权重，模型加入了加性注意力机制[20]，计算过程如式(2)—(3)所示：

$$c_{i,k}=\boldsymbol{U}_{\text{hidden}}^{\text{T}}\tanh(\boldsymbol{W}_{\text{hidden}}\boldsymbol{h}_{i,k}^{\text{T}}+\boldsymbol{b}_{\text{hidden}}) \tag{2}$$

$$\alpha_{i,k}^{\text{text}}=\frac{\exp(c_{i,k})}{\sum_{j=0}^{n}\exp(c_{i,k})} \tag{3}$$

其中：$c_{i,k}$是模型计算得到的注意力分数，表示 $\boldsymbol{h}_{i,k}$在句子中所占的情感权重。权重矩阵 $\boldsymbol{W}_{\text{hidden}}$和偏置向量 $\boldsymbol{b}_{\text{hidden}}$用于将 $\boldsymbol{h}_{i,k}$映射到注意空间，然后将投影与上下文向量 $\boldsymbol{U}_{\text{hidden}}$(随机初始化并在训练中学习)相乘。将 $c_{i,k}$通过 Softmax 标准化，产生情感权重 $\alpha_{i,k}^{\text{text}}$。将单词特征按对应的情感权重进行加权，得到最终的文本情感特征 $\boldsymbol{T}_{\text{i}}^{\text{final}}$，如式(4)所示：

$$\boldsymbol{T}_{i}^{\text{final}}=\sum_{k=0}^{n}\alpha_{i,k}^{\text{text}}\times\boldsymbol{h}_{i,k} \tag{4}$$

1.3 多层图像特征提取

相比于从图像中检测目标、识别物体，理解图像的情感往往更加困难，其关键点在于对图像的不同抽象层次进行有效地提取。在 CNN 模型中，不同位置与尺寸的卷积核，其对图像特征的提取有着不同的偏好，Zeiler 等[21-22]对卷积神经网络的隐藏层

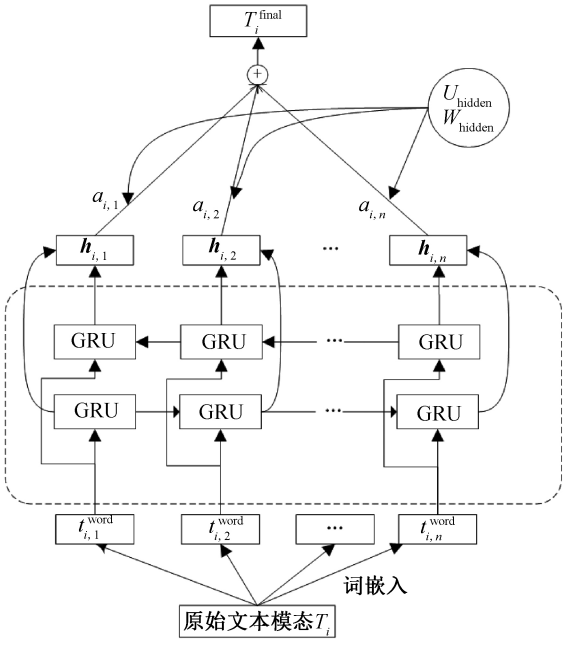

图 2 文本特征提取模块

特征进行反卷积可视化，证明了不同层次卷积核所提取的特征类型是不同的，即：低层卷积核所学习到的往往是颜色、边缘等低层特征，中间层卷积通常会学习到网格、纹理等特征，而高层卷积学习到的则是完整、具有辨别性的高层语义。在目标检测中，为了检测出图像的类型，只需要利用最高层卷积特征。但在情感分析中，图像的情感往往隐藏在中、低层的颜色、纹理等特征中，仅用高层卷积难以有效识别图像情感。因此，本文以 VGG13 网络为原型，构建了如图 3 所示的多层图像特征提取网络。多层图像特征提取网络由 10 个卷积层与 5 个最大池化层组成，每经过一个池化层，外接一层卷积以提取该层特征，将图像特征的提取从 VGG 网络的最后一层扩展为了多层。该模型不仅能获取图像高层语义，还能有效提取图像美学分析与图像色彩等中低层特征。

完整地训练一个深度卷积神经网络需要一个足够大的数据集和大量的训练时间，由于实验的数据集都是中等大小，且硬件设备有限，所以本文采用迁移学习的方法，将在 ImageNet 数据集预训练好的 VGG 网络作为基准模型。将模型新增加的外接卷积层参数随机初始化，其余层则使用在 ImageNet 上预训练的 VGG13 模型的权重。在训练时，对模型进行微调，冻结了原 VGG13 网络层参数，只将外接卷积层与前馈层参与到模型训练中。

给定一个图文对($\boldsymbol{P}_i$，$\boldsymbol{T}_i$)，对于图像模态 $\boldsymbol{P}_i$，将其转换成维度为(224，224，3)的矩阵，输入至多层

特征提取网络，会得到由高到低 5 个不同层次的图像特征，其特征表示为 $\boldsymbol{P}_{i,k}\in\mathbf{R}^{h\times w\times f}$，其中 w 代表图像的宽度，h 代表图像的高度，f 代表特征通道数。然后重塑图像特征的形状，将 w 与 h 合并到同一维度，并得到新的图像区域特征表示 $\boldsymbol{P}_{i,k}\in\mathbf{R}^{S\times f}$，其中 $S=w\times h$，模块结构如图 3 所示。

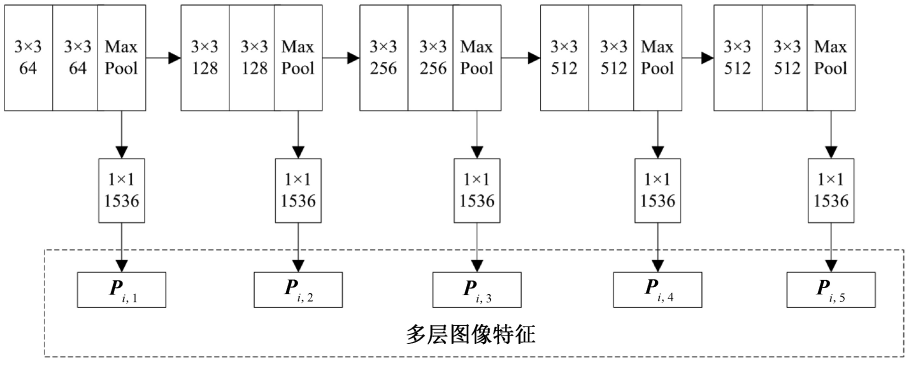

图 3　多层图像特征提取模块

1.4　跨模态特征融合

多模态情感分析的核心在于不同模态特征的提取、跨模态特征融合以及融合后特征的分类。本模块聚焦于前期融合（特征层融合），在单模态情感特征提取后，融合各单模态模型的情感特征，模块结构如图 4 所示。本文结合已建立的单模态特征提取网络，即 Bi-GRU 模型和 VGG13 模型，分别提取文本和图像的情感特征，为了挖掘图像与文本之间的情感共现，学习模态之间的相关性与互补性，我们采用跨模态注意力公式将文本特征与多层图像特征进行融合。跨模态注意力公式包含查询（Query）、键（Key）和值（Value）三个基本元素，通过将查询项与键进行相关性计算，得到每个键所对应值的权重值，并通过 Softmax 函数与缩放点乘对这些权重进行归一化，最后将权重和相应的值进行加权求和得到最后的融合特征。将文本特征与训练矩阵 $\boldsymbol{W}_{\text{query}}$ 相乘，作为查询矩阵，各层图像特征分别与训练矩阵 $\boldsymbol{W}_{\text{key}}$、$\boldsymbol{W}_{\text{value}}$ 相乘后，作为键矩阵和值矩阵，将上述矩阵通过注意力公式，得到单层文本－图像注意力特征 $\boldsymbol{A}_{i,k}$，计算公式如式(5)—(6)所示：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

$$\boldsymbol{A}_{i,k}=\text{Attention}(\boldsymbol{T}_i^{\text{final}}\boldsymbol{W}_{\text{query}},\boldsymbol{P}_{i,k}\boldsymbol{W}_{\text{key}},\boldsymbol{P}_{i,k}\boldsymbol{W}_{\text{value}}) \tag{6}$$

不同层次的文本－图像注意力特征的情感信息都是不一样的，对于情感特征蕴含在图像中某一实体的情况，高层文本－图像注意力特征在情感检测方面往往更有意义，而当情感特征蕴含在图像色彩、构图中时，低层次的文本－图像注意力特征应该分配更高的权重。对此，本文再次使用加性注意力，给不同层次的单层文本－图像注意力特征分配权重。通过一层非线性激活函数 tanh 的神经元，将第 k 层图文特征的表征 $\boldsymbol{A}_{i,k}$ 通过 $\boldsymbol{W}_{\text{attention}}$ 与 $\boldsymbol{b}_{\text{attention}}$ 矩阵投射到注意空间。然后将投影与上下文向量 $\boldsymbol{U}$（随机初始化并在训练中学习）相乘，得到标量 $u_{i,k}$，表示第 k 层图文特征的相对重要性。将 $u_{i,k}$ 通过 Softmax 标准化，产生权重 $\alpha_{i,k}$。最后，$\boldsymbol{S}_i$ 向量表示是由其所有层表示 $\boldsymbol{A}_{i,k}$ 及其注意权重 $\alpha_{i,k}$ 的加权求和产生的。计算流程如式(7)—(9)所示：

$$u_{i,k}=\boldsymbol{U}^{\mathrm{T}}\tanh(\boldsymbol{W}_{\text{attention}}\boldsymbol{A}_{i,k}^{\mathrm{T}}+\boldsymbol{b}_{\text{attention}}) \tag{7}$$

$$\alpha_{i,k}=\frac{\exp(u_{i,k})}{\sum_{j=1}^{5}\exp(u_{i,j})} \tag{8}$$

$$\boldsymbol{s}_i=\sum_{j=1}^{5}\alpha_{i,k}\times\boldsymbol{A}_{i,k} \tag{9}$$

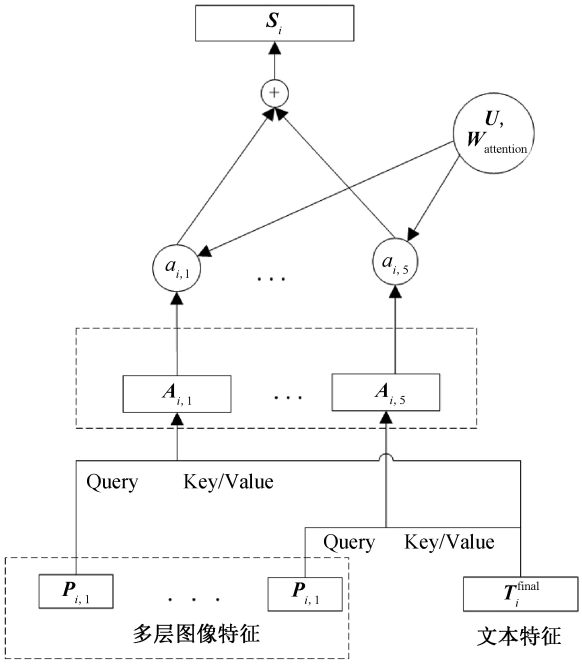

图 4　跨模态特征融合模块

1.5 情感分类输出

在将图文特征进行融合后,将其输入到多层感知机中,并通过 Softmax 函数进行情感分类。多层感知机由多个前馈层组成,将 Cross_entropy 作为损失函数,对模型进行训练:

$$loss=\frac{1}{D}\sum_{k=1}^{d}y'_{i,k}\log(y_{i,k}) \tag{10}$$

其中:$loss$ 表示损失值,D 为图文评论样本总数,d 为情感类别数,$y_{i,k}$ 为第 i 个图文对预测为第 k 类情感的概率,$y'_{i,k}$ 为指示变量,若与真实样本类别相同,则 $y'_{i,k}$ 为 1,不同则为 0。

2 实验与结果分析

2.1 数据集

本文在对于 Memotion-7K 和 MVSA 数据集上进行仿真实验。对于 Memotion-7K 数据集是国际语义评测大会为参赛者研究图文情感分析所提供的公开数据集,该数据集收录了近些年流行的表情包与对应的文字,共包含 6992 条图文样本,情感标签分为“非常积极”、“积极”、“中性”、“消极”和“非常消极”五类。MVSA 数据集是图文情感分析领域常用的数据集,其样本是从 Twitter 上收集的图文评论,由于参与注释的人数不同,该数据集包含 MVSA-Multi 和 MVSA-Single 两个版本,本文将 MVSA-Multi 数据集中注释不统一的数据筛除,并与 MVSA-Single 数据集进行合并,其情感标签分为“积极”、“中性”和“消极”三类,数据集样本统计如表 1 所示。

表 1 数据集统计

数据集	样本总数/条	有效样本数/条	类型数/类	不同类型样本/条				
				非常消极	消极	中性	积极	非常积极
MVSA	25521	22481	3	—	6488	7531	8462	—
Memotion-7K	6992	6204	5	151	480	2201	2339	1033

对于 Memotion-7K 数据集,其“非常消极”与“消极”类别的样本数要明显少于其他类别,属于不平衡分类样本集,为了防止模型产生偏见,在实验中对“非常消极”与“消极”两类样本进行过采样,并增加其在计算损失函数时的损失权重。

2.2 数据预处理

在实验开始之前,对数据进行了预处理。对于 MVSA 数据集,因为推特文本中通常会频繁出现“@用户名”、“#标签”和“网址链接”等难以识别且对情感分析无意义的内容,其存在会影响训练模型的准确率,因此将单词中以“@”、“#”和“http”开头的单词删除,例如,原句“How I feel today #legday #jelly #aching #gym”经过清洗后变为“How I feel today”。对于 Memotion-7K 数据集,其文本是从图像中转化而来的,而其中部分图像添加了网站的水印,所以本文将文本中出现的网址都进行删除,例如“TORTILLA CHIPS MARVEL CAPTAIN AMERICA CIVIL WAR imgflip.com”经过处理后变为“TORTILLA CHIPS MARVEL CAPTAIN AMERICA CIVIL WAR”。利用 Python 中的 Torchvision 包读取图像,对图像进行处理,将其大小调整为 224×224×3,适应 VGG13 网络的输入。

2.3 评价指标

对于 MVSA 数据集和 Memotion-7K 数据集,本文将其按顺序分为训练集(80%)、验证集(10%)和测试集(10%),采用准确率(Accuracy)和 F1 值(F1-Score)对情感分类的性能进行评估,F1 值是综合考虑了精确率和召回率的指标,计算公式如式(11)—(12)所示:

$$P_{\mathrm{A}}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}+N_{\mathrm{FP}}+N_{\mathrm{TN}}} \tag{11}$$

$$\begin{cases}P_{\mathrm{R}}=\dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\\ P_{\mathrm{P}}=\dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\\ P_{\mathrm{F}}=\dfrac{2\times P_{\mathrm{R}}\times P_{\mathrm{P}}}{P_{\mathrm{R}}+P_{\mathrm{P}}}\end{cases} \tag{12}$$

其中:N_{TP}是正确地被标记为正例的样本数,N_{FP}是被错误地标记为正例但实际上是反例的样本数,N_{TN}是被正确地标记为反例的样本数,N_{FN}是被错误地标记为反例但实际上是正例的样本数,P_{A} 表示准确率,P_{R} 表示召回率,P_{P} 表示准确率,P_{F} 表示 F1 值。

2.4 实验参数设置

本文使用 Pytorch 框架作为模型的定义、训练与测试平台,在训练与测试阶段使用 2 个 RTX3090 作为硬件平台。对于文本特征提取方面,采用“BERT-Base, Uncase”预训练模型进行动态词嵌入,将句子最大长度设置为 40,词嵌入维度为 768,Bi-GRU 模块的隐藏层维度为 768。实验采用的损失函数是 Cross_entropy,Mini-batch 大小设置为

64，学习率统一设置为 0.0003，优化器采用 Adam[23]，在每个全连接层之后采用 Dropout，Dropout 率设为 0.1。在模型训练中，采用 Early-stoping 技术，训练轮数设置为 100 轮，Patience 设置为 8，当验证集损失值连续 8 次都没有下降，训练就停止，可以有效防止过拟合[24]。

2.5　消融实验

为了验证 MCAF 模型的有效性，本文进行了两组对比实验：一是将 MCAF 模型与其变体对比，二是同情感分析基线模型对比。本文提出的 MCAF 模型变体与基线模型如下：

a) MCAF-1 模型。该模型仅将最高层图像特征与文本特征进行注意力融合，得到图像-文本注意力融合特征后输入多层感知机进行情感分类。

b) MCAF-2 模型。该模型将多层图像特征分别与文本特征进行注意力融合，得到多组单层图像-文本注意力融合特征，将多组融合特征相加，再输入多层感知机进行情感分类。

c) MCAF 模型。该模型将多层图像特征分别与文本特征进行注意力融合，得到多组单层图像—文本注意力融合特征，将多组融合特征通过加性注意力重新分配权重后相加，再输入多层感知机进行情感分类。

d) VGG13 模型。VGG13 是经典的图像分类模型，本文将经过 Imagenet 预训练之后的 vgg13 网络进行微调，用于图像情感分类。

e) Bi-GRU 模型。Bi-GRU[25] 采用传统 Glove 词嵌入与双向 GRU 网络对的文本情感进行分类。

f) Early Fusion。该模型将 VGG13 模型获取的图像特征和 Bi-GRU 模型获取的文本特征进行拼接，并输入多层感知机进行情感分类。

g) Late Fusion。Late Fusion 模型将 VGG13 和 Bi-GRU 模型通过后期融合，将两个模型的情绪得分进行加权与分类。

2.6　定量分析

表 2 显示了基线模型、本文所提 MCAF 模型及其变体在 MVSA 数据集和 Memotion-7K 数据集上的准确率和 F1 值对比。总体来说，本文提出的模型在两个实验数据集上的表现都要优于其变体与基线模型。早期融合模型性能要优于单模态的情感分析模型，因为多模态数据包含了更多原始特征，模型接受的数据更加完整准确，特征融合后对状态的提取更加充分，对于单模态数据，不论是文本还是图像，在预测多模态情感时都丢失了另一模态的信息，导致难以预测整体情感。对比 MCAF-1 模型与早期融合模型，在 MVSA 数据集上，MCAF 模型比早期融合模型在分类准确率上提升了 0.49%，在 F1 值上提升了 0.96%；在 Memotion-7K 数据集上，MCAF-1 比早期融合模型在分类准确率上提升了 1.21%，在 F1 上提升了 0.89%。MCAF-1 的性能有所提升，是因为与早期融合模型直接将图像情感特征与文本情感特征拼接的特征融合方法相比，MCAF-1 使用注意力机制融合两个模态的特征，令模态间相关的情感特征获得更高的权重，能更有效的结合多模态信息，从而提高情感分类的性能。

表 2　不同情感分析模型的性能对比

模型	算法	MVSA 数据集		Memotion-7K 数据集	
		准确率	F1 值	准确率	F1 值
基准模型	VGG13	0.5655	0.5971	0.4345	0.4211
	Bi-GRU	0.6610	0.6424	0.6078	0.5918
	Early Fusion	0.6729	0.6428	0.6224	0.6134
	Late Fusion	0.6303	0.6167	0.5837	0.5822
本文模型	MCAF-1	0.6778	0.6524	0.6345	0.6223
	MCAF-2	0.6557	0.6310	0.5921	0.5830
	MCAF	0.6990	0.6784	0.6549	0.6497

对比 MCAF 模型与 MCAF-1 模型，在 MVSA 数据集上，MCAF 模型比 MCAF-1 模型在分类准确率上提升了 2.12%，在 F1 值上提升了 2.60%；在 Memotion-7K 数据集上，MCAF 比 MCAF-1 模型在分类准确率上提升了 2.04%，在 F1 值上提升了 2.74%。MCAF 模型的性能要高于 MCAF-1 模型，这是因为相比只利用卷积网络的最后一层输出，提取图像的多层特征可以更完整的获取图像蕴含的情感，其与文本融合后的情感特征也更加全面。MCAF 模型与 MCAF-2 模型都采用多层图像特征与文本特征进行注意力融合的方法，区别在于 MCAF 使用加性注意力对融合特征再次加权，而 MCAF-2 将权重平均分配，在 MVSA 和 Memotion-7K 两个数据集上，MCAF 比 MCAF-2 提高

4.33%、6.28%的准确率与 4.74%、6.67%的 F1 值,对于每个样本,其不同图像层次所蕴含的情感特征是不平等的,不同层次的图文情感特征的贡献程度会对最后的情感分类结果产生影响,MCAF 模型通过软注意力机制分配不同层次图像情感特征的权重,会比平均分配权重产生更好的效果,令重要程度更高的图文融合情感特征被输入到分类层。此外,MCAF-2 模型的性能要明显低于 MCAF-1 模型,原因是在本文实验数据集中,多数图文样例的情感蕴含在高层语义中,其最高层的图文情感特征会占据更高的权重,导致 MCAF-2 模型的性能偏低。

2.7 定性分析

为了更清晰地探究多层图像特征提取与注意力融合对多模态情感分析所起到的作用,本节将从“多层图像特征提取”和“注意力融合”两个方面对 MCAF 模型进行定性分析,图 5 为从测试集中选取的不同类型的图像样例,表 3 给出了选取样例的文本信息与早期融合模型、MCAF-1 和 MCAF 模型对不同类型图文样例的预测结果。

(a) 样例1 (b) 样例2 (c) 样例3

(d) 样例4 (e) 样例5 (f) 样例6

图 5 测试集中的图像样例

表 3 不同模型对测试集样例的预测结果

图像内容	文本内容	真实标签	早期融合模型	MCAF-1	MCAF
样例 1	Partying with cousins!	积极	积极	积极	积极
样例 2	Honestly, the largest, and most delicious, slice of chocolate cake I′ve ever seen.	积极	中性	积极	积极
样例 3	Thick smoke in the air.	消极	消极	中性	消极
样例 4	The pre-show folks trying to be funny makes my face go like this.	消极	消极	消极	消极
样例 5	NationalDogDay to the dog I miss every single day??	消极	积极	中性	中性
样例 6	In Calgary and marvelling at these western skies. I missed this. Great to be back.	积极	中性	中性	积极

早期融合模型、MCAF-1 和 MCAF 模型使用相同的文本特征提取网络,区别在于早期融合模型使用拼接进行特征融合,MCAF-1 使用注意力融合代替拼接,MCAF 模型在 MCAF-1 的基础上额外提取了多层图像特征。通过将上述模型的分类结果进行对比,可以有效地展现“注意力融合”和“多层图像特征提取”对情感分类性能的提升。早期融合模型由于难以检测图文之间的情感共现,仅对样例 1 与样例 4 这类含有明显情感标识(如微笑、皱眉),且文本情感与图像情感相一致的样例有较高的识别率,对于实体类或风景类图文样例,该模型往往难以识别;MCAF-1 由于使用了注意力融合,令文本与图像中相关联的特征获取更高的权重,其与早期融合模型相比,在样例 2 与样例 5 这类文本内容与图像实体相关的样例有更高的识别率;而 MCAF 模型通过将图像的不同层特征与文本特征进行注意力融合,实

现了文本特征与不同层次图像特征的交互，不仅能准确识别样例 1、样例 2 与样例 4，在识别样例 3、样例 6 这类风景类图文样例时，也能通过利用图像的色彩、构图，联合文本信息理解图像情感。以上模型对测试集图文样例的识别结果，充分证明了多层跨模态注意力融合模型在不同类型图文情感分析中的有效性。

3　结　论

在社交网络中，越来越多的用户将文本与各种类型的图像结合来表达情感，其中图像的情感往往隐藏在不同特征层中，而目前大多情感分析模型忽略了图像的低层次特征与文本特征的关联，未能有效利用低层图像特征。针对这一问题，本文提出了基于多层跨模态注意力融合的图文情感分析模型，该模型首先使用 Bi-GRU 充分利用了上下文话语间的互补信息，其次使用改进 VGG13 网络以提取多层图像情感特征；通过注意力融合将两个模态之间的交互信息进行了有效地挖掘，考虑了不同模态之间的联系；最后使用加性注意力为每层图文融合特征分配权重，使权重较大的特征层能更好地决定情感分类结果。实验结果显示，本文模型比传统的多模态情感分析模型在准确率和 F1 值上均有明显提升，表明多层跨模态注意力融合能有效地将图像各层特征与文本特征相结合。

由于从社交媒体中采集的文本中包含许多俚语和缩写单词，并不包含在词嵌入模型的词表中，将其处理为未知单词会影响模型的性能。因此后续研究可尝试建立词典，将未记录在词表的高频俚语和缩写单词映射成常规单词，以降低在文本模态存在的噪音、提升情感分析模型性能。

参考文献：

[1] 林敏鸿，蒙祖强. 基于注意力神经网络的多模态情感分析[J].计算机科学，2020，47(S2)：508-514.

[2] Pérez-Rosas V，Mihalcea R，Morency L P. Utterance-level multimodal sentiment analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia，Bulgaria：ACL，2013：973-982.

[3] 范涛，吴鹏，曹琪. 基于深度学习的多模态融合网民情感识别研究[J]. 信息资源管理学报，2020，10(1)：39-48.

[4] Poria S，Chaturvedi I，Cambria E，et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]// 2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona，Spain：IEEE，2016：439-448.

[5] Cao D L，Ji R R，Lin D Z，et al. A cross-media public sentiment analysis system for microblog[J]. Multimedia Systems，2016，22(4)：479-486.

[6] 缪裕青，汪俊宏，刘同来，等. 图文融合的微博情感分析方法[J]. 计算机工程与设计，2019，40(4)：1099-1105.

[7] 凌海彬，缪裕青，张万桢，等. 多特征融合的图文微博情感分析[J]. 计算机应用研究，2020，37(7)：1935-1939.

[8] Zhao Z Y，Zhu H Y，Xue Z H，et al. An image-text consistency driven multimodal sentiment analysis approach for social media[J]. Information Processing & Management，2019，56(6)：97-102.

[9] 谢豪，毛进，李纲. 基于多层语义融合的图文信息情感分类研究[J].数据分析与知识发现，2021，5(6)：103-114.

[10] 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. Copenhagen，Denmark. Stroudsburg，PA，USA：Association for Computational Linguistics，2017：1103-1114.

[11] 郭可心，张宇翔.基于多层次空间注意力的图文评论情感分析方法[J/OL].计算机应用：1-9.（2021-01-26）[2021-05-23]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210125.1709.007.html.

[12] Li M G，Li W R，Wang F，et al. Applying BERT to analyze investor sentiment in stock market[J]. Neural Computing and Applications，2021，33(10)：4663-4676.

[13] Chen J Y，Yan S K，Wong K C. Verbal aggression detection on Twitter comments：convolutional neural network for short-text sentiment analysis[J]. Neural Computing and Applications，2020，32(15)：10809-10818.

[14] Campos V，Jou B，Giró-I-Nieto X. From pixels to sentiment：Fine-tuning CNNs for visual sentiment prediction[J]. Image and Vision Computing，2017，65：15-22.

[15] Zhao S，Gao Y，Jiang X，et al. Exploring principles-of-art features for image emotion recognition [C]// Proceedings of the 22nd ACM international conference on Multimedia. Orlando，Florida，USA：ACM Press，2014：47-56.

[16] Gu J，Wang Z，Kuen J，et al. Recent advances in convolutional neural networks[J]. Pattern Recognition，2018，77：354-377.

[17] Rao T, Li X, Xu M. Learning multi-level deep representations for image emotion classification[J]. Neural Processing Letters, 2020, 51(3): 2043-2061.

[18] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language under standing[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. 2019: 4171-4186.

[19] Tang D Y, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 1422-1432.

[20] Truong Q T, Lauw H W. VistaNet: Visual aspect attention network for multimodal sentiment analysis [J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 305-312.

[21] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2014: 818-833.

[22] Zeiler M D, Taylor G W, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning[C]//2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011: 2018-2025.

[23] Kingma D P, Ba J L. Adam: A method for stochastic optimization[C]//Proceedings of the 2015 International Conference on Learning Representations. San Diego: ICLR, 2015: 1-13.

[24] Sakurai Y, Yoshikawa M, Faloutsos C. FTW: Fast similarity search under the time warping distance[C]// Proceedings of the Twenty Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York: ACM Press: 2005: 326-337.

[25] Sachin S, Tripathi A, Mahajan N, et al. Sentiment analysis using gated recurrent neural networks[J]. SN Computer Science, 2020, 1(2): 1-13.

(责任编辑:康 锋)