计算机科学
COMPUTER SCIENCE

基于多尺度跨模态特征融合的图文情感分类模型

刘倩, 白志豪, 程春玲, 归耀城

引用本文

刘倩, 白志豪, 程春玲, 归耀城. 基于多尺度跨模态特征融合的图文情感分类模型[J]. 计算机科学, 2024, 51(9): 258-264.

LIU Qian, BAI Zhihao, CHENG Chunling, GUI Yaocheng. Image-Text Sentiment Classification Model Based on Multi-scale Cross-modal Feature Fusion [J]. Computer Science, 2024, 51(9): 258-264.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700163

基于多尺度跨模态特征融合的图文情感分类模型

刘 倩[1] 白志豪[1] 程春玲[1] 归耀城[2]
1 南京邮电大学计算机学院、软件学院、网络空间安全学院 南京 210023
2 南京邮电大学现代邮政学院 南京 210023
(qianliu@njupt.edu.cn)

摘 要 图文情感分类任务常用早期融合和Transformer模型相结合的跨模态特征融合策略进行图文特征融合,但该策略更倾向于关注模态内部的独有信息,而忽略了模态间的相互联系和共有信息,导致跨模态特征融合效果不理想。针对此问题,提出一种基于多尺度跨模态特征融合的图文情感分类方法。局部尺度方面,基于跨模态注意力机制进行局部特征融合,使模型不仅关注图像和文本的独有信息,而且可以发现图像和文本之间的联系和共有信息。全局尺度方面,基于MLM损失进行全局特征融合,使模型对图像和文本数据进行全局建模,进一步挖掘图像和文本之间的联系,从而促进图像和文本特征的深度融合。在两个公开数据集MVSA-Single和MVSA-Multiple上与10个基线模型进行对比实验,结果表明所提方法在精度、F1值和模型参数量方面均具有明显优势,验证了其有效性。

关键词:图文情感分类;跨模态特征融合;Transformer模型;注意力机制;MLM损失
中图分类号 TP391.1

Image-Text Sentiment Classification Model Based on Multi-scale Cross-modal Feature Fusion

LIU Qian[1],BAI Zhihao[1],CHENG Chunling[1] and GUI Yaocheng[2]
1 School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 School of Modern Posts,Nanjing University of Posts and Telecommunications,Nanjing 210023,China

Abstract For the image-text sentiment classification task,the cross-modal feature fusion strategy which combines early fusion and Transformer model is usually used for image-text feature fusion. However,this strategy prefers to focus on the unique information within a single modality,while ignoring the interconnections and common information among multiple modalities,resulting in unsatisfactory effect of cross-modal feature fusion. To solve this problem,a method of image-text classification based on multi-scale cross-modal feature fusion is proposed. On the one hand,for the local scale,local feature fusion is carried out based on the cross-modal attention mechanism,so that the model not only focuses onthe unique information of the image and text,but also explores the connection and common information between the image and text. On the other hand,for the global scale,global feature fusion based on MLM loss enables the model to conduct global modeling of image and text data,further mine the relationship between them,and thus promote the deep fusion of image and text features. Compared with ten baseline models on two public datasets,MVSA-Single and MVSA-Multiple,the proposed method shows distinct advantages in accuracy,F1 score,and model parameter quantity,verifying its effectiveness.

Keywords Image-Text sentiment classification,Cross-modal feature fusion,Transformer model,Attention mechanism,MLM loss

1 引言

随着使用富文本编辑器的社交媒体快速发展,社交平台的内容形式不再局限于单一的文字,越来越多携带个人情感和观点的图像、语音、视频等多模态评论数据大量涌现。这些评论在虚拟世界中不断传播和发酵,进而影响现实世界中事态的发展。从社交平台的多模态数据中识别和分析用户情感已经成为当下情感分析领域的研究热点[1-2]。根据调查显示,社交平台用户更倾向于选择图像和文本这两种数据来传递情感信息[3],因此本文重点研究对图文数据的情感分类,即基于图文数据预测用户的积极、中性、消极3种情感类别。

跨模态特征融合是多模态任务的核心内容,而去除模态的噪声[4]和冗余信息,发现模态间的交互关系和模态内部的独有信息是实现跨模态特征融合的关键。图文情感分类任务常用的早期融合的跨模态特征融合策略将图像特征和文本特征进行拼接作为图文特征融合的结果。尽管该方法相较于

到稿日期:2023-07-21 返修日期:2023-12-28
基金项目:江苏省双创博士项目(JSSCBS20210507)
This work was supported by the Foundation of Jiangsu Provincial Double-Innovation Doctor Program(JSSCBS20210507).
通信作者:程春玲(chengcl@njupt.edu.cn)

单模态模型在性能上有所提升，但由于图像和文本具有异构性和异质性[5]，简单的融合策略无法充分挖掘模态间的共有信息和模态内部的独有信息，限制了模型性能。近年来，Transformer[6]网络的快速发展使得多模态特征在同一模型中进行交互成为现实。越来越多图文领域的研究者采用Transformer网络的自注意力机制实现图像和文本特征之间的交互以解决早期融合方法存在的问题。但Transformer网络的自注意力机制更倾向于关注模态内部的独有信息，忽略了模态间的关联和共有信息，导致跨模态特征融合效果并不理想。

针对此问题，Xu等[7]考虑了图像和文本的关联，在特征拼接操作之前对图像和文本进行特征交互，并验证了其有效性。受该方法启发，本文提出了一种基于跨模态特征融合的图文情感分类方法，从局部和全局两个尺度帮助模型建立图像和文本的联系。一方面，基于跨模态注意力机制进行局部特征融合。与自注意力机制寻找同一模态数据的独有信息不同，跨模态注意力机制能够较好地发现图文之间的关联信息[8-9]。将基于跨模态注意力机制获得的图文关联信息和原始的图像、文本信息一起输入Transformer网络，通过其自注意力机制的动态加权操作进一步发现图文之间的联系以及图文的独有信息，从而获得更好的特征融合效果。另一方面，基于MLM(Masked Language Modeling)损失进行全局特征融合。考虑到MLM具有强大的全局建模能力，模型加入MLM损失作为辅助损失函数，有利于从全局进一步挖掘图像和文本之间的关联，促进图文特征的深度融合。与传统MLM损失进行随机掩码的方式不同，本文方法每次遮盖与图像相关的文本内容以减少噪声对模型训练的影响。

综上所述，本文的贡献如下：

1)提出了一种基于跨模态注意力的局部特征融合方法，既关注图文数据的独有信息又关注其联系和共有信息；

2)提出了一种基于MLM损失的全局特征融合方法，通过对图文数据的全局理解引导模型进一步挖掘图文数据的关联；

3)提出了一种改进MLM损失掩码过程的方法，每次仅遮盖与图像相关的文本取代随机掩码，减少噪声对模型训练的影响。

2　相关工作

图文情感分类的早期探索为图像模态和文本模态提取了合适的特征表示。近年来，多数工作使用CNN网络或者其变体来提取图像特征，使用BERT[10]模型来提取文本特征，在这些工作的帮助下图文情感分类的性能得到了快速提升。

2.1　多模态特征融合与注意力机制

根据多模态特征融合阶段，可将多模态特征融合划分为早期融合、中期融合和晚期融合[11]。早期融合被应用在许多图文情感分类模型中。Cai等[12]预先训练文本CNN和图像CNN获取文本和图像的表示，随后将这些特征向量拼接后输入全连接层进行分类；Xu等[13]观察到图像中的物体信息和场景信息会与情感产生联系，使用VGG[14]网络提取图像中的场景信息和物体信息，使用LSTM[15]网络提取文本特征信息，将多种特征拼接后输入全连接层进行情感分类；Cheema等[16]对图像进行细粒度的特征提取，将得到的细粒度的多种特征拼接融合后用于提升多模态融合特征的表达能力。这些方案在下游数据集中取得了较好的性能。

使用简单拼接操作的早期融合方法无法充分挖掘图文特征间的联系。随着注意力机制的快速发展，基于注意力机制的特征能够有针对性地关注重要内容，这一优势使其可用于学习图像和文本之间的关联。Xu等[7]提出了一种记忆网络，利用多层堆叠的双流网络结构在每一层中使用注意力机制建模图像和文本之间的联系，最终将得到的图文全局特征拼接后送入全连接层进行分类；Li等[17]将拼接后的图文融合特征送入Transformer网络，通过自注意力机制执行图文特征融合，挖掘图文之间的深层次联系；Wei等[18]提出TGF模块(Text-Guided Fusion Module)来解决视觉部分存在冗余特征的问题，并使用对比学习解决表征空间的特征移位问题；Wang等[19]采用CNN和Transformer网络相结合的方式，将拼接后的图文融合特征送入CNN网络以捕获局部细节，然后使用Transformer网络捕获远距离特征依赖。此外，图卷积神经网络(Graph Convolutional Networks，GCN)因具有能够捕获局部和全局的数据结构模式的特点[20]，被广泛应用在图文情感分类领域中。

除早期融合方法外，还有一些中期和晚期融合方法被应用于图文情感分类领域。中期融合方法中，Liao等[21]和Yang等[22]将GCN与注意力机制相结合，在图文情感分类任务中取得了优异性能；Jiang等[23]提出交互信息融合模块，旨在捕获图文模态表征之间的相互作用，并提出了特定信息提取模块用于提取更加丰富的图文特征；Peng等[24]提出特征注意力模块用于捕获图文模态表征之间的相互作用，引入Image-Caption模型将其生成的图像描述文本送入该特征注意力模块以生成更好、更准确的图像注意力特征。晚期融合方法中，Yu等[25]分别针对图像和文本做情感分类，随后根据两种模态的情感分类结果使用平均策略得到最终的结果。

2.2　MLM损失

在模型中加入MLM损失的目的是训练语言模型，使其更好地理解文本的上下文，从而根据学习到的信息预测文本中遮盖的词。MLM损失具有较好地建模上下文关系的优点，因此被研究者应用于多模态领域中对图像和文本的语义联系进行建模。Li等[26]使用ITM(Image Text Matching)和MLM损失预训练Transformer结构的网络模型，在多数任务中取得了最优性能；Zhao等[27]使用MLM损失作为其中一个训练任务，模型在下游的多模态数据集上取得了优异的性能；Sun等[28]将BERT模型应用于视频模态中，使用BERT模型自带的MLM损失和NSP(Next Sentence Prediction)损失作为其中两个训练损失，在下游数据集上实现了最优性能。

Transformer网络强大的自注意力交互能力使得早期融合加Transformer网络的结构成为当前图文情感分类领域的主流模型之一，在下游任务中取得了较好的性能。但自注意力机制在运行过程中会将图像和文本模态视为单一模态进行处理，导致模型无法有效获取图像和文本的独有信息及其关

联信息，影响图文特征融合效果。受注意力机制和 MLM 损失的启发，本文提出了一种基于跨模态注意力和 MLM 损失的图文情感分类模型，用于解决 Transformer 网络的自注意力机制无法很好地促进图文特征融合的问题。

3 模型架构

本文提出的基于多尺度跨模态特征融合的图文情感分类模型整体结构如图 1 所示，包括特征提取层、局部特征融合层和全局特征融合层 3 部分。

特征提取层负责将图像和文本数据分别转化为其对应的特征向量；局部特征融合层基于跨模态注意力机制进行局部尺度的特征融合，用于发现图像和文本之间的联系和共有信息，以及图像和文本的独有信息；全局特征融合层基于 MLM 损失进行全局尺度的特征融合，用于进一步发现图像和文本之间的联系和共有信息。全局特征融合层和局部特征融合层共享参数一方面是为了减少模型参数量，加速训练和推理；另一方面是为了更好地促进局部特征融合层和全局特征融合层进行信息交流以获得更准确的情感分类结果。

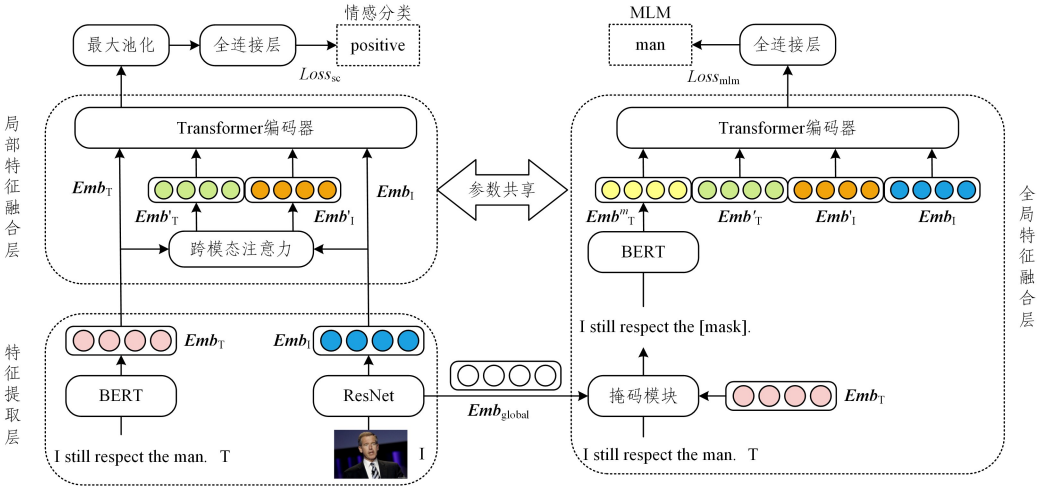

图 1 基于多尺度跨模态特征融合的图文情感分类模型的整体结构

Fig. 1 Overall structure of image-text sentiment classification model based on multi-scale cross-modal feature fusion

3.1 特征提取

给定一个图像文本对(I,T)，其中图像数据 $I\in\mathbb{R}^{H\times W\times C}$，文本数据 $T=\{t_1,t_2,\cdots,t_n\}$ 包含 n 个词汇，特征提取层分别将图像和文本数据转化为其对应的特征向量。

3.1.1 图像特征提取

ResNet 模型[29]的残差连接较好地解决了梯度消失问题，网络模型可以获取更好的特征提取效果，因此本文使用 ResNet 模型提取图像特征，过程如下：

$$\boldsymbol{Emb}_{\text{local}}=ResNet(I;\theta_{\text{pre}}^{\text{ResNet}}) \tag{1}$$

$$\boldsymbol{Emb}_{\text{global}}=AvgPooling(\boldsymbol{Emb}_{\text{local}}) \tag{2}$$

$$\boldsymbol{E}_I=Concat(\boldsymbol{Emb}_{\text{global}},\boldsymbol{Emb}_{\text{local}}) \tag{3}$$

$$\boldsymbol{Emb}_I=\text{gelu}(\boldsymbol{W}_I\boldsymbol{E}_I+\boldsymbol{b}_I) \tag{4}$$

其中，$\boldsymbol{Emb}_{\text{local}}\in\mathbb{R}^{m*d_i}$ 为图像局部特征向量，$\boldsymbol{Emb}_{\text{global}}\in\mathbb{R}^{1*d_i}$ 为图像全局特征向量，$\theta_{\text{pre}}^{\text{ResNet}}$ 为 ResNet 模型的预训练权重，$\boldsymbol{Emb}_I\in\mathbb{R}^{d_t}$ 为最终表示图像数据的特征向量，$\boldsymbol{W}_I\in\mathbb{R}^{d_t\times d_i}$ 为可训练的权重矩阵，$\boldsymbol{b}_I\in\mathbb{R}^{d_t}$ 为可训练的偏差项，$AvgPooling$ 表示池化操作，$Concat$ 表示特征拼接操作，gelu 为激活函数。

3.1.2 文本特征提取

BERT 模型在 NLP 领域的许多下游任务中取得了先进性能，因此本文用 BERT 模型对文本进行编码，过程如下：

$$\boldsymbol{E}_T=BERT(T;\theta_{\text{pre}}^{\text{BERT}}) \tag{5}$$

$$\boldsymbol{Emb}_T=\text{gelu}(\boldsymbol{W}_T\boldsymbol{E}_T+\boldsymbol{b}_T) \tag{6}$$

其中，$\theta_{\text{pre}}^{\text{BERT}}$ 为 BERT 模型使用的预训练权重，$\boldsymbol{W}_T\in\mathbb{R}^{d_t\times d_t}$ 为可训练的权重矩阵，$\boldsymbol{b}_T\in\mathbb{R}^{d_t}$ 为可训练的偏差项，$\boldsymbol{Emb}_T\in\mathbb{R}^{n*d_t}$ 为表示文本数据的特征向量。

3.2 局部特征融合

跨模态注意力机制[8-9]可以更好地发现多模态数据间的关联信息。Jiang 等[23]在跨模态处理中使用全连接层寻找源模态数据与经过跨模态处理的数据之间的联系。受该工作启发，本文也使用跨模态注意力机制，将文本的每一个元素和图像所有元素做注意力操作得到图像特征引导的文本特征向量矩阵 $\boldsymbol{Emb}_T'$，将图像的每一个元素和文本所有元素做注意力操作得到文本特征指导的图像特征向量矩阵 $\boldsymbol{Emb}_I'$，这样做的目的是能够较好地发现图像数据和文本数据之间的细粒度关联信息。但与 Jiang 等[23]不同的是，本文将寻找源模态数据与经过跨模态处理的数据之间的联系交由特征抽取能力更强大的 Transformer 模型处理。

本文的跨模态注意力模块如图 2 所示，$\odot$表示内积操作，Ⓣ表示转置操作。跨模态注意力计算过程如下：

$$\boldsymbol{Att\_s\_m}=\boldsymbol{Emb}_T Trans(\boldsymbol{Emb}_I) \tag{7}$$

$$\boldsymbol{Att\_w\_m}_{I\to T}=\text{softmax}(\boldsymbol{Att\_s\_mEmb}) \tag{8}$$

$$\boldsymbol{Emb}_T'=\boldsymbol{Att\_w\_m}_{I\to T}\odot\boldsymbol{Emb}_I \tag{9}$$

$$\boldsymbol{Att\_w\_m}_{T\to I}=\text{softmax}(Trans(\boldsymbol{Att\_s\_m})) \tag{10}$$

$$\boldsymbol{Emb}_I'=\boldsymbol{Att\_w\_m}_{T\to I}\boldsymbol{Emb}_T \tag{11}$$

其中，$\boldsymbol{Att\_s\_m}\in\mathbb{R}^{n*(1+m)}$ 为注意力分数矩阵，$\boldsymbol{Att\_w\_mEmb}_{I\to T}\in\mathbb{R}^{n*(1+m)}$ 为衡量文本局部特征和不同图像局部特征关联程度的权重矩阵，$\boldsymbol{Att\_w\_m}_{T\to I}\in\mathbb{R}(1+m)*n$ 为衡量图像局部特征和不同文本局部特征关联程度的权重矩阵，$Trans$ 表示转置操作，$\odot$表示内积操作。最终获得的 $\boldsymbol{Emb}_T'\in\mathbb{R}^{n*d_t}$ 和

$\boldsymbol{Emb}_I' \in \mathbb{R}^{(1+m)*d_t}$ 保存了图像和文本数据的关联和共有信息。

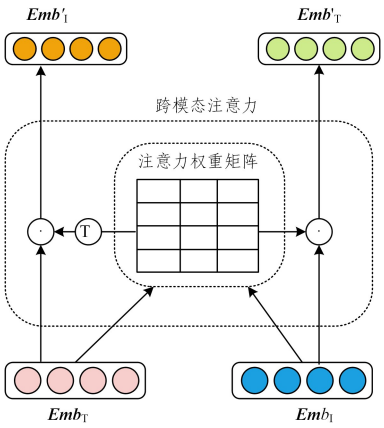

图 2　跨模态注意力模块结构

Fig. 2　Structure of cross-modal attention module

基于跨模态注意力机制的局部特征融合模块随后将 $\boldsymbol{Emb}_T'$ 和 $\boldsymbol{Emb}_I'$ 作为辅助信息与原始的图像特征向量 $\boldsymbol{Emb}_I$ 和文本特征向量 $\boldsymbol{Emb}_T$ 一起作为 Transformer 编码器的输入，Transformer 编码器的自注意力机制在辅助信息的帮助下通过动态加权的方式可以进一步发现图像和文本之间的联系和共有信息，以及图像和文本的独有信息，从而获得更好的特征融合效果。图像和文本局部特征融合过程如下：

$$\boldsymbol{Emb}_f = T(Concat(\boldsymbol{Emb}_T, \boldsymbol{Emb}_T', \boldsymbol{Emb}_I', \boldsymbol{Emb}_I)) \tag{12}$$

其中，$\boldsymbol{Emb}_f \in \mathbb{R}^{d_t}$ 为局部特征融合后得到的图文特征融合向量，T 为 Transformer 网络。

3.3　全局特征融合

MLM 任务具有强大的全局建模能力，该任务在输入文本中随机遮盖一些词语并要求模型预测其原始值，帮助模型加强对文本上下文的理解[10]。本文在图文情感分类任务常用的情感分类损失 $Loss_{sc}$ 基础上，加入 MLM 任务的损失函数 $Loss_{mlm}$ 作为辅助损失函数，通过要求模型根据剩余的文本内容和图像信息预测遮盖的文本内容的方式使模型对图文数据进行全局建模，进一步挖掘图像和文本之间的关联，从而促进图文特征的深度融合。模型的最终训练损失为：

$$Loss_{total} = Loss_{sc} + Loss_{mlm} \tag{13}$$

3.3.1　情感分类损失

情感分类损失是图文情感分类模型中常用的损失，计算过程如下：

$$\boldsymbol{Emb}_{maxPool} = MaxPooling(\boldsymbol{Emb}_f) \tag{14}$$

$$L_{pre} = \mathrm{softmax}(\boldsymbol{W}_p \boldsymbol{Emb}_{maxPool} + \boldsymbol{b}_p) \tag{15}$$

$$Loss_{sc} = CrossEntropy(L_{pre}, L_{true}) \tag{16}$$

其中，$MaxPooling$ 为池化操作，$CrossEntropy$ 为交叉熵操作，$\boldsymbol{W}_p \in \mathbb{R}^{3*d_t}$ 为可训练的权重矩阵，$\boldsymbol{b}_p \in \mathbb{R}^3$ 为可训练的偏差项，L_{pre} 为模型预测的情感标签，L_{true} 为真实情感标签。

3.3.2　MLM 损失

文本中可能包含许多噪声，如 URL 链接、与内容无关的广告和不合理的标点符号等。为避免模型遮盖大量噪声文本起到反效果，本文提出掩码模块对原始 MLM 损失的掩码过程进行改进，不是进行随机掩码，而是遮盖与图像最相关的文本，从而减少噪声对模型训练的影响。

掩码模块结构如图 3 所示，Ⓜ表示取最大值操作，Ⓡ表示替换操作。掩码模块计算过程如下：

$$T_{tokenized} = BertTokenizer(T) \tag{17}$$

$$\boldsymbol{S\_m} = \mathrm{softmax}(\boldsymbol{Emb}_{global} \odot Trans(\boldsymbol{Emb}_T)) \tag{18}$$

$$in\_set = \max(\boldsymbol{S\_m}, n = mlm * len(\boldsymbol{Emb}_T)) \tag{19}$$

$$T_{mask} = Replace(T_{tokenized}, ref = in\_set) \tag{20}$$

$$\boldsymbol{Emb}_T^m = BERT(T_{mask}) \tag{21}$$

其中，$BertTokenizer$ 为 Bert 模型分词器，$T_{tokenized}$ 为分词后的文本数据，$\boldsymbol{S\_m} \in \mathbb{R}^{1*n}$ 为衡量图像信息和文本信息的相似度矩阵，in_set 为要遮盖的文本数据的下标集合，mlm 为掩码率，$\max(A, n=num)$ 表示从 A 中取出前 num 个值最大的元素下标，$Replace$ 为替换操作，$Replace(A, ref=in\_set)$ 表示将 in_set 中下标在 A 中对应的元素替换为“[mask]”，$\boldsymbol{Emb}_T^m \in \mathbb{R}^{n*d_t}$ 为带有“[mask]”标记的文本特征向量。

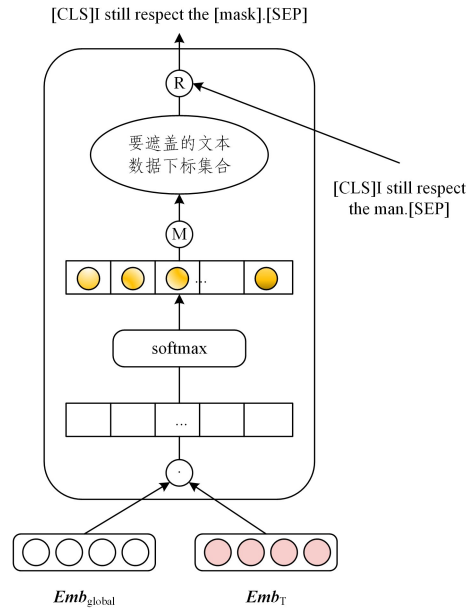

图 3　掩码模块结构

Fig. 3　Structure of masking module

基于改进掩码过程的 MLM 损失计算过程如下：

$$\boldsymbol{Emb}_f' = T(Concat(\boldsymbol{Emb}_T^m, \boldsymbol{Emb}_T', \boldsymbol{Emb}_I', \boldsymbol{Emb}_I)) \tag{22}$$

$$\boldsymbol{Emb}'_{text} = cut(\boldsymbol{Emb}_f', n = len(\boldsymbol{Emb}_T^m)) \tag{23}$$

$$L_{p\_token} = \mathrm{softmax}(\boldsymbol{W}_q \boldsymbol{Emb}'_{text} + \boldsymbol{b}_q) \tag{24}$$

$$Loss_{mlm} = CrossEntropy(L_{p\_token}, L_{t\_token}) \tag{25}$$

其中，$\boldsymbol{Emb}_f' \in \mathbb{R}^{(2n+2m+2)*d_t}$ 为全局特征融合后的图文特征融合向量，$cut(a, n)$ 表示截取操作，参数 n 表示截取的长度，$\boldsymbol{Emb}'_{text} \in \mathbb{R}^{n*d_t}$ 为图文特征融合向量中文本的向量表示，L_{p_token} 为模型在“[mask]”标记处预测的文本标签，$\boldsymbol{W}_q \in \mathbb{R}^{d_t*d_{voc}}$ 为可训练的权重矩阵，$\boldsymbol{b}_q \in \mathbb{R}^{d_{voc}}$ 为可训练的偏差项，d_{voc} 为 Bert 模型词表大小，L_{t_token} 为“[mask]”标记处真实的文本标签。

4　实验与分析

本章详细介绍了实验数据集，说明了实验的相关设置，并对比和分析了本文方法与 10 个基线方法的实验结果。此外通过可视化的方法进一步分析了本文方法的有效性。

4.1 数据集

MVSA-Single 和 MVSA-Multiple 数据集[30]（简称为 MVSA 数据集）是图文情感分类任务常用的数据集。为了和其他方法进行公平比较，本文采用 Xu 等[13]的方法对 MVSA 数据集进行处理，按照 8∶1∶1 的比例分割为训练集、验证集和测试集。MVSA 数据集的详细信息如表 1 所列。

表 1 MVSA 数据集信息

Table 1 Details of MVSA dataset

数据集	Train	Val	Test	合计
MVSA-Single	3611	450	450	4511
MVSA-Multiple	13624	1700	1700	17024

4.2 实验设置

使用 HuggingFace Transformer[31]和 Pytorch 来实现本文模型，使用 Bert-base-uncased 和 ResNet50 作为文本编码器和图像编码器。对于 MVSA-Single 和 MVSA-Mulitple 数据集，batch_size 分别设置为 64，64，学习率分别设置为 2×10^{-5}、4×10^{-5}，Transformer encoder 层数分别设置为 3 和 4，掩码率都采用 25%，使用 AdamW 优化器进行模型参数的学习。硬件配置为 Intel(R) Xeon(R) Platinum 8350C 处理器和 A5000 显卡，内存 32 GB，显存 24 GB。模型训练过程中固定 Bert-base-uncased 模型和 ResNet50 模型的参数，更新其他参数。

4.3 对比模型

本文所提模型的特征融合策略属于早期融合，因此选择 6 个使用早期融合策略的多模态图文情感分类模型进行对比。对比模型如下：

CNN-Multi[12]：基于 CNN 的模型。

MultiSentiNet[13]：基于深层语义网络的模型。

Co-MN-Hop[7]：基于协同记忆网络的模型。

CLMLF[17]：基于对比学习和多层融合的模型。

MVCN[18]：基于多视图校准网络解决模态异质性的模型。

CLCAF[19]：基于对比学习和密集注意力的模型。

此外，为进一步验证所提模型的有效性，本文选择 1 个使用晚期融合策略和 3 个使用中期融合策略的方法进行对比。对比模型如下：

DNN-LR[25]：基于深度卷积神经网络的模型。

FENet[23]：基于特征融合网络的模型。

ITIGNN[21]：基于图神经网络的模型。

CMCN[24]：基于层次融合的跨模态互补网络的模型。

4.4 实验结果及分析

4.4.1 与基线模型的比较

本文选择精度(Acc)、F1 值和模型参数量作为评价指标。表 2 列出了基线模型和本文所提模型的实验结果。从表 2 可以看出，本文模型在 MVSA-Single 数据集上的性能优于其他所有基线模型。在 MVSA-Multiple 数据集上，与早期融合方法相比，本文模型取得了和最好的 MVCN 模型相当的性能，同时使用的模型参数量比 MVCN 模型减少了 31%，比 CLMLF 模型减少了 56%；与晚期融合和中期融合方法相比，本文模型的精度优于所有基线模型，F1 值略差于 FENet 模型和 CMCN 模型，模型参数量相比 FENet 模型和 CMCN 模型分别减少了 29%和 27%。总体上本文模型和其他基线模型相比展现出了明显优势。

表 2 不同模型在 MVSA 数据集上的实验结果

Table 2 Results of different models on MVSA dataset

模型	MVSA-Single		MVSA-Multiple		参数量
	Acc/%	*F1*/%	*Acc*/%	*F1*/%	
DNN-LR[2016]	61.42	61.03	67.86	66.33	4.43×10^6 *
FENet[2020]	74.21	74.06	71.46	**71.21**	115.78×10^6 *
ITIGNN[2022]	73.84	74.04	—	—	55.97×10^6 *
CMCN[2022]	73.61	75.03	70.45	70.45	113.60×10^6 *
CNN-Multi[2015]	61.20	58.37	66.39	64.19	4.32×10^6 *
MultiSentiNet[2017]	69.84	69.63	68.86	68.11	5.83×10^6 *
Co-MN-Hop[2018]	70.51	70.01	69.92	69.83	4.22×10^6 *
CLMLF[2022]	75.33	73.46	72.00	69.83	187.35×10^6
MVCN[2023]	76.06	74.55	**72.07**	70.01	120.12×10^6 *
CLCAF[2023]	76.44	75.61	70.53	67.45	80.86×10^6 *
本文模型‡	74.67	73.78	71.12	66.89	129.42×10^6
本文模型	**76.89**	**75.83**	72.06	69.37	82.65×10^6

注：* 表示估计值，由于无法获取模型的源代码，因而给出估计值供参考；
‡ 表示不采用共享参数策略所训练的模型。

分析表 2 可知，CNN-Multi 用深度卷积神经网络获得图文特征向量，经过简单拼接后预测情感分类，该方法忽略了图文特征之间的关联。MultiSentiNet 在深度网络提取的特征基础上，考虑了图文之间的单向关联，模型性能较 CNN-Multi 有所提升。Co-MN-Hop 对图像到文本和文本到图像两个方向的关联进行建模，进一步提升了模型性能。CLMLF 使用 BERT 获得文本特征，使用 Transformer 模型的自注意力机制进行图文特征融合而不是简单的特征拼接，又使用对比学习促进特征交互，因此在 MVSA-Multiple 上其 F1 值达到最优，精度为次优。MVCN 对 CLMLF 模型结构进行修改，提出 TGF 模块以解决视觉信息存在冗余特征的问题，并使用不同于 CLMLF 的对比学习解决表征空间的特征移位问题，因而其性能优于 CLMLF。CLCAF 在 CLMLF 的 Transformer 网络之前添加 CNN 网络以捕捉图文特征的局部信息，并用对比学习拉远不同类别特征的距离，因此在 MVSA-Single 上其精度和 F1 值均超越了 CLMLF。本文模型分两个尺度分别进行图文特征的局部融合和全局融合，能够发现模态间的共有信息和模态内部的独有信息，因而能够获得整体最优性能。与不使用共享参数策略的“本文模型‡”相比，“本文模型”在减少参数量的同时，又能够更好地协调局部和全局特征融合层进行信息交流，因此取得了更好的性能。

4.4.2 掩码率实验

由于社交平台的文本较短，较低的掩码率会导致遮盖的文本信息太少，不利于模型训练；而较高的掩码率会导致模型无法学习到足够的上下文信息，影响模型对图文数据的整体理解。为此本小节进行了不同掩码率的实验，找到最适合本文模型的掩码率。模型使用不同掩码率在图文情感分类任务上的实验结果如表 3 所列。实验结果表明，最适合本模型的掩码率为 25%。

表 3　MLM 损失中不同掩码率的实验结果

Table 3　Results at different mask rates in MLM loss

（%）

掩码率	MVSA-Single		MVSA-Multiple	
	Acc	*F*1	*Acc*	*F*1
15%	75.35	74.90	71.23	68.55
20%	75.78	75.22	71.64	69.26
25%	**76.89**	**75.83**	**72.06**	**69.37**
30%	76.22	75.03	71.47	69.02

4.4.3　消融实验

本小节进行消融实验，分析本文模型各个部分的独立作用。消融实验的结果如表 4 所列。“-MLM 损失”表示去掉基于 MLM 损失的全局特征融合，只使用基于跨模态注意力的局部特征融合；“-跨模态注意力”表示局部特征融合时仅使用图像和文本的独有特征，去掉跨模态的特征交互，加上基于 MLM 损失的全局特征融合；“-MLM 损失-跨模态注意力”表示同时去掉 MLM 损失和跨模态注意力。

从表 4 的实验结果可知，加上跨模态注意力和 MLM 损失的图文情感分类模型（本文模型）优于其他模型，证明两种方法相结合有助于模型实现最佳性能。从表 4 还可以观察到，与只使用跨模态注意力相比（-MLM 损失），只使用 MLM 损失的模型（-跨模态注意力）性能更好，表明从全局的角度对图文数据建模更有助于模型提升性能。我们猜测这是因为文本数据中包含较多噪声，如 URL 链接、大量的“#”符号等，所以使用细粒度的跨模态注意力进行局部特征建模会使模型容易受到噪声的影响，从而降低在测试集上的性能；而使用 MLM 损失的模型是对整体的图文内容进行建模，使得模型不用过多地关注细粒度的局部噪声；此外，改进掩码过程的掩码模块可以进一步避免噪声的影响，因此使用 MLM 损失的模型性能更好。仅使用图像和文本的独有信息进行特征融合而不考虑图文关联和共有信息的模型（-MLM 损失-跨模态注意力）在所有模型中性能最差，验证了模态间的相互联系和共有信息对于跨模态特征融合的重要性。

表 4　消融实验结果

Table 4　Ablation results

（%）

模型	MVSA-Single		MVSA-Multiple	
	Acc	*F*1	*Acc*	*F*1
本文模型	**76.89**	**75.83**	**72.06**	**69.37**
-MLM 损失	75.33	74.52	70.82	66.18
-跨模态注意力	75.56	74.63	71.64	67.72
-MLM 损失-跨模态注意力	74.44	72.33	70.58	67.19

4.4.4　可视化实验

为了更直观地观察本文模型的优势，我们对 CLMLF 模型和本文模型进行了可视化对比分析，结果如表 5 所列，图片数据中的非浅蓝色区域表示模型在做出情感决策时所关注的区域，浅绿色、黄色、红色 3 种颜色表示模型关注相应区域的程度由低到高逐渐增加；图像和文本数据中红色的颜色越深，表明模型在做出情感预测时越关注此部分信息。相较于“+跨模态注意力”的结果，“+MLM 损失”的结果更能够关注到与情感相关的图文本内容，因此“+MLM 损失”相较于“+跨模态注意力”的性能更好，这与消融实验中的分析结果一致；“+跨模态注意力+MLM 损失（本文模型）”使模型既关注到了图像中与情感相关的部分，又关注到了文本中与情感相关的部分，且与 CLMLF 模型相比，本文模型更能够关注到和情感相关的图文内容，所以本文模型的性能更好。

表 5　本文模型和 CLMLF 模型的案例分析可视化结果（电子版为彩图）

Table 5　Visualized results of case study of the proposed model and CLMLF model

原始数据	+跨模态注意力	+MLM 损失	本文模型	CLMLF 模型
				
A great night to meet the residents in Deer Run!	A great night to meet the residents in Deer Run!	A great night to meet the residents in Deer Run!	A great night to meet the residents in Deer Run!	A great night to meet the residents in Deer Run!
				
Since he's already rejecting the premise and he's always rejecting questions.	Since he's already rejecting the premise and he's always rejecting questions.	Since he's already rejecting the premise and he's always rejecting questions.	Since he's already rejecting the premise and he's always rejecting questions.	Since he's already rejecting the premise and he's always rejecting questions.

结束语　本文提出了一种基于多尺度跨模态特征融合的图文情感分类方法，分别从局部和全局两个角度帮助模型同时发现图像和文本之间的关联和共有信息以及图像和文本的独有信息，从而更好地进行图文特征融合。本文方法在 MVSA-Single 和 MVSA-Multiple 数据集上的实验结果优于一系列对比基线模型，验证了该方法的有效性。

MVSA-Single 和 MVSA-Multiple 数据集中存在多语种数据，目前本文只是简单地将这些多语种数据和英文数据一同输入模型进行情感分类，这种做法使模型性能受到影响。在未来工作中，我们将重点研究多语种数据的情感分析，提升模型对多语种图文数据的情感分类能力。

参 考 文 献

[1]　ZHANG L，WANG S，LIU B. Deep learning for sentiment analysis：A survey[J]. Wiley Interdisciplinary Reviews：Data Mining and Knowledge Discovery，2018，8(4)：e1253.

[2] GUO Y X, JIN Y, TANG H, et al. Multi-modal Emotion Recognition Based on Dynamic Convolution and Residual Gating[J]. Computer Engineering, 2023, 49(7): 94-101.
[3] AN X. Research on image-text sentiment analysis method based on cross-modal fusion [D]. Beijing: Beijing University of Technology, 2020.
[4] PETZ G, KARPPWICZ M, FURSCHUß H, et al. Reprint of: Computational approaches for mining user's opinions on the Web 2. 0[J]. Information Processing & Management, 2015, 51(4): 510-519.
[5] BALTRUSAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.
[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of NIPS'17. 2017: 6000-6010.
[7] XU N, MAO W, CHEN G. A co-memory network for multimodal sentiment analysis[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 929-932.
[8] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C] // Proceedings of CVPR'17. 2017: 299-307.
[9] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]// Proceedings of ECCV'18. 2018: 201-216.
[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of NAACL'19. 2019: 4171-4186.
[11] AI-TAMEEMI I K S, FEIZI-DERAKHSHI M R, PASHAZADEH S, et al. A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks[J]. arXiv: 2207.02160, 2022.
[12] CAI G, XIA B. Convolutional neural networks for multimedia sentiment analysis [C]// Proceedings of NLPCC'15. 2015: 159-167.
[13] XU N, MAO W. Multisentinet: A deep semantic network for multimodal sentiment analysis [C]// Proceedings of CIKM'17. 2017: 2399-2402.
[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.
[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[16] CHEEMA G S, HAKIMOV S, MULLER-BUDACK E, et al. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods[C]// Proceedings of MMPT'21. 2021: 37-45.
[17] LI Z, XU B, ZHU C, et al. CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection [C]// Findings of NAACL'22. 2022: 2282-2294.
[18] WEI Y, YUAN S, YANG R, et al. Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection[C] // Proceedings of ACL'23. 2023: 5240-5252.
[19] WANG H, LI X, REN Z, et al. Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion[J]. Sensors, 2023, 23(5): 2679.
[20] ZHANG Z, CUI P, ZHU W. Deep learning on graphs: A survey [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 249-270.
[21] LIAO W, ZENG B, LIU J, et al. Image-text interaction graph neural network for image-text sentiment analysis[J]. Applied Intelligence, 2022, 52(10): 11184-11198.
[22] YANG X, FENG S, ZHANG Y, et al. Multimodal sentiment detection based on multi-channel graph neural networks[C]// Proceedings of ACL'21. 2021: 328-339.
[23] JIANG T, WANG J, LIU Z, et al. Fusion-extraction network for multimodal sentiment analysis[C]// Proceedings of PAKDD'20. 2020: 785-797.
[24] PENG C, ZHANG C, XUE X, et al. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification[J]. Tsinghua Science and Technology, 2021, 27(4): 664-679.
[25] 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.
[26] LI J, 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.
[27] ZHAO J, LI R, JIN Q, et al. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition[C]// Proceedings of ICASSP'22. 2022: 4703-4707.
[28] SUN C, MYERS A, VONDRICK C, et al. Videobert: A joint model for video and language representation learning[C]// Proceedings of ICCV'19. 2019: 7464-7473.
[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of CVPR'16. 2016: 770-778.
[30] NIU T, ZHU S, PANG L, et al. Sentiment analysis on multiview social data[C]// Proceedings of MMM'16. 2016: 15-27.
[31] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-art natural language processing[C]// Proceedings of EMNLP'20. 2020: 38-45.

LIU Qian, born in 1986, Ph.D, lecturer, is a member of CCF(No. 98989M). Her main research interests include artificial intelligence and sentiment analysis.

CHENG Chunling, born in 1972, professor, is a member of CCF (No. E200015597M). Her main research interests include data mining and data management.

(责任编辑:何杨)