计算机科学
COMPUTER SCIENCE

基于文本和图像门控融合机制的多模态方面级情感分析

张添植，周刚，刘洪波，刘铄，陈静

引用本文

张添植，周刚，刘洪波，刘铄，陈静. 基于文本和图像门控融合机制的多模态方面级情感分析[J]. 计算机科学，2024，51(9)：242-249.

ZHANG Tianzhi, ZHOU Gang, LIU Hongbo, LIU Shuo, CHEN Jing. Text-Image Gated Fusion Mechanism for Multimodal Aspect-based Sentiment Analysis [J]. Computer Science, 2024, 51(9): 242-249.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

面向物联网的分布式联邦学习加密验证研究

Study on Cryptographic Verification of Distributed Federated Learning for Internet of Things

计算机科学, 2024, 51(6A): 230700217-5. https://doi.org/10.11896/jsjkx.230700217

基于对比学习的视觉增强多模态命名实体识别

Vision-enhanced Multimodal Named Entity Recognition Based on Contrastive Learning

计算机科学, 2024, 51(6): 198-205. https://doi.org/10.11896/jsjkx.230400052

基于深度哈希学习的知识库问答检索框架

Deep Hashing-based Retrieval Framework for KBQA

计算机科学, 2023, 50(11): 227-233. https://doi.org/10.11896/jsjkx.220900206

基于增强序列标注策略的单阶段联合实体关系抽取方法

Single-stage Joint Entity and Relation Extraction Method Based on Enhanced Sequence Annotation Strategy

计算机科学, 2023, 50(8): 184-192. https://doi.org/10.11896/jsjkx.220700082

增强实体表示的文档级关系抽取方法研究

Study on Enhanced Entity Representation for Document-level Relation Extraction

计算机科学, 2023, 50(8): 157-162. https://doi.org/10.11896/jsjkx.220700161

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600117

基于文本和图像门控融合机制的多模态方面级情感分析

张添植[1]　周　刚[1,2]　刘洪波[1]　刘　铄[1]　陈　静[1]
1 战略支援部队信息工程大学　郑州 450001
2 数学工程与先进计算国家重点实验室　郑州 450001
(timothyz2023@163.com)

摘　要　多模态方面级情感分析是多模态情感分析领域的一项新兴任务,旨在对给定的方面实体在文本和图像中所体现的情感进行识别。尽管多模态方面级情感分析研究近年来取得了突破性的进展,但是现有的模型在多模态特征融合阶段大都仅采用简单的拼接方法,而没有考虑图像中是否存在与文本语义不相关的信息,这在一定程度上可能会为模型引入额外的噪声。为了解决上述问题,提出了一种基于文本和图像门控融合机制的多模态方面级情感分析模型(TIGFM)。该模型在文本和图像进行交互的同时引入了从数据集图像中提取的形容词-名词对(ANPs),并将其中形容词的加权作为图像辅助信息;此外,在特征融合阶段,通过构建一种动态控制图像和图像辅助信息输入的门控机制实现多模态特征融合。实验结果表明,TIGFM 模型在两个基于 Twitter 的数据集上取得了具有竞争力的结果,进而验证了所提方法的有效性。

关键词:多模态方面级情感分析;门控融合机制;形容词-名词对;图像辅助信息;语义相关性
中图分类号　TP391

Text-Image Gated Fusion Mechanism for Multimodal Aspect-based Sentiment Analysis

ZHANG Tianzhi[1],ZHOU Gang[1,2],LIU Hongbo[1],LIU Shuo[1] and CHEN Jing[1]
1 Information Engineering University,Zhengzhou 450001,China
2 State Key Laboratory of Mathematical Engineering and Advanced Computing,Zhengzhou 450001,China

Abstract　Multimodal aspect-based sentiment analysis is an emerging task in multimodal sentiment analysis field,which aims to identify the sentiment of each given aspect in text and image. Although recent research on multimodal sentiment analysis has made breakthrough progress,most existing models only use simple concatenation in multimodal feature fusion without considering whether there is semantically irrelevant information in image with text,which may introduce additional interference to the model. To address the above problems,this paper proposes a text-image gated gusion mechanism(TIGFM) model for multimodal aspect-based sentiment analysis,which introduces adjective-noun pairs(ANPs) extracted from the dataset images while text interacts with image,and treats the weighted adjectives as image auxiliary information. In addition,multimodal feature fusion is achieved by constructing a gating mechanism that dynamically controls the input of image and image auxiliary information in feature fusion stage. Experimental results demonstrate that TIGFM model achieves competitive results on two Twitter datasets,and then validate the effectiveness of proposed method.

Keywords　Multimodal aspect-based sentiment analysis,Gated fusion mechanism,Adjective-noun pairs,Image auxiliary information,Semantic relevance

1　引言

随着互联网技术的飞速发展,人们逐渐开始参与网络上的各类活动并在线分享自身的观点和看法。目前,社交媒体和众多服务平台在人们的日常生活中广泛普及,同时,用户针对不同话题和事件逐渐倾向于发布文本和图像等相结合的多模态信息来表达自身的情感,这一趋势吸引了学术界对多模态情感分析研究的广泛关注。近年来,多模态情感分析已成为情感计算领域的研究热点[1],其在政治选举、股市预测以及医疗保健[2]等实际工作中具有极大的应用价值。

多模态方面级情感分析(Multimodal Aspect-Based Sentiment Analysis,MABSA)是多模态情感分析研究中一项重要的细粒度任务,其通过选定文本中的某些名词作为方面实体并结合文本和图像内容进一步推断该方面实体的情感极性。图 1 展示了一个具有代表性的示例,其中方面实体为文本中的“Klay Thompson”,可以通过文本外加图像信息的辅助

到稿日期:2023-06-14　返修日期:2024-02-23
基金项目:河南省科技攻关项目(222102210081)
This work was supported by the Science and Technology Research Program of Henan Province(222102210081).
通信作者:周刚(gzhougzhou@126.com)

进而预测出方面实体的情感极性为消极。相较于多模态全局情感分析，多模态方面级情感分析是一项更加精细、更具挑战性的情感分析任务，其可以捕获全局情感分析无法获取的文本内部实体的情感极性。

[Klay Thompson]$_{Neg}$ looking like he already planning what he gonna do after they get eliminated… # NBAPlayoffs

图 1 多模态方面级情感分析的代表性示例

Fig. 1 Representative example of MABSA

鉴于该领域的重要性和较为广泛的应用前景，当前研究人员已提出众多的多模态方面级情感分析模型。例如，MIMN[3]和 ESAFN[4]通过注意力机制对方面实体、文本和图像实现模态间的交互，TomBERT[5]和 Saliencybert[6]使用预先训练的语言和视觉模型对文本单词特征和图像视觉特征实现有效的编码以及更好的表示学习等。这些研究成果证明了将图像融合至传统的文本情感分析中能够使模型获得更为准确的情感预测能力。然而，这些模型在多模态特征融合阶段大都仅采用简单的拼接方法构建最终的特征表示进而实现情感分析，但是它们并未考虑图像中存在与文本语义不相关的信息可能会为模型引入额外的噪声。

为了解决这一问题，本文提出了一种基于文本和图像门控融合机制的多模态方面级情感分析模型（Text-Image Gated Fusion Mechanism for Multimodal Aspect-Based Sentiment Analysis，TIGFM）。针对数据集中的某一样本，该模型首先使用预训练的文本和图像编码器分别获得文本和图像的特征表示，然后通过跨模态注意力机制实现文本和图像的交互获得方面实体感知的图像表示。此外，模型引入从数据集图像中所提取的形容词-名词对（ANPs）[7]，并将其中形容词的加权作为图像辅助信息来增强图像语义的表达能力。最后，通过设计一种多模态特征融合门控机制动态控制图像信息对最终特征表示的贡献程度，进而对样本的情感极性进行预测。

本文工作的主要贡献如下：

（1）提出了一种基于文本和图像门控融合机制的多模态方面级情感分析模型 TIGFM，该模型在文本和图像进行交互的同时引入了从图像中所提取的 ANPs，并将其中形容词的加权作为图像辅助信息，进而获得对图像更好的语义表示和情感表达；

（2）在获得文本、图像以及图像辅助信息表示之后，模型在最后的特征融合阶段通过构建一种多模态特征融合门控机制动态控制图像和图像辅助信息表示的输入，防止其中与文本语义不相关的信息为模型引入额外的噪声；

（3）相较于大多数现有的多模态方面级情感分析模型，TIGFM 模型在两个基于 Twitter 的数据集上取得了最佳的性能效果。

2 相关工作

在早期的情感分析研究中，主要是针对文本级和图像级等单模态情感分析进行研究。近年来，多模态情感分析逐渐成为该研究领域中的关注重点，多模态方面级情感分析也在传统方面级情感分析的研究基础上进一步得到发展与完善。

2.1 多模态情感分析

当前，多模态情感分析研究在学术界受到广泛关注[8]，其通过将文本和其他非文本模态（如视觉和听觉模态等）信息相结合实现模型的构建，并主要围绕会话和社交媒体数据两项子任务展开研究。在针对会话的多模态情感分析研究中，现有的方法主要侧重于采用不同的深度学习模型（如卷积神经网络、长短期记忆网络、门控循环单元和 Transformer 等）对不同模态的信息进行交互，目前已被证明在诸多研究任务中有着较好的表现（如情感分析[9-10]、情绪分析[11-12]和讽刺检测[13-14]等）；在针对社交媒体的多模态情感分析中，主要的工作聚焦于针对社交媒体图像的情感分析[15-16]以及针对图像和文本相结合的多模态情感分析[17-18]。然而，上述方法主要适用于粗粒度情感分析（即识别每个样本中所体现的全局情感）研究，并不能直接在多模态方面级的细粒度情感分析任务中使用。

2.2 方面级情感分析

作为一项重要的细粒度情感分析研究，在过去的十年中方面级情感分析在自然语言处理领域有着较为广泛的研究与应用[19]，其所使用的方法大致可分为基于离散特征和基于深度学习的两类。基于离散特征的方法通过设计多种指定方面的特征训练用于情感分析的学习分类器[20]；基于深度学习的方法采用各类神经网络模型（如递归神经网络[21]、卷积神经网络[22]、循环神经网络[23]、注意力机制[24-25]、图卷积神经网络[26-27]以及预训练语言模型 BERT[28-29]等）对方面实体和其所对应上下文进行编码。然而，上述方法仅适用于单独的文本模态，并未考虑其他模态的相关信息也会为情感分析研究做出一定的贡献。

2.3 多模态方面级情感分析

为了有效利用不同模态的信息，实现方面级情感分析研究，近三年来众多学者使用不同任务中的多种有效方法构建了一系列的多模态方面级情感分析模型。Xu 等[3]首先对该研究任务进行探索并提出了一种基于多交互双向长短期记忆网络的模型 MIMN，实现文本和图像的交互，同时构建了一个基于电商评论的数据集 ZOL；Yu 等[5]提出了一种基于 BERT 架构改进的模型 TomBERT，同时也构建了两个基于 Twitter 的数据集 Twitter-2015 和 Twitter-2017；Yu 等[4]提出了一种基于实体敏感注意力和融合网络的模型 ESAFN；Khan 等[30]提出了一种基于跨模态翻译将图像内容转换为文本字幕且仅通过文本实现最终情感分析的模型 CapBERT。然而，上述方法虽然在多模态方面级情感分析研究中被证明具有一定的有效性，但是它们往往未考虑图像中可能存在与文本语义不相关的信息，这在一定程度上会为模型引入额外的噪声。为此，本文提出的模型通过构建一种多模态特征融合门控机制来有效地解决这一问题。

3 模型架构

TIGFM 模型的整体架构如图 2 所示。该模型共包含 4 个模块：(1)特征提取模块；(2)特征交互模块；(3)特征辅助模块；(4)特征融合模块。本章首先对多模态方面级情感分析研究任务进行定义，然后进一步详细介绍 TIGFM 模型中的每个模块。

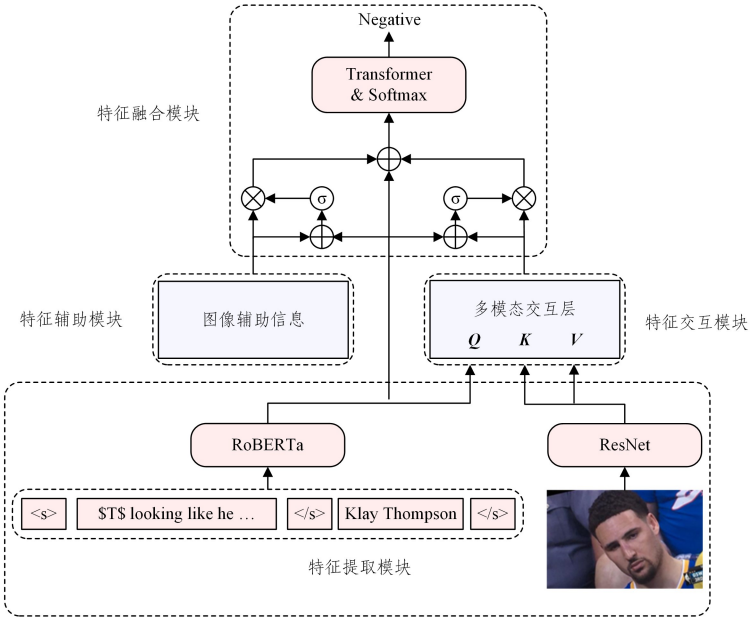


图 2 TIGFM 模型的整体架构

Fig. 2 Overall architecture of TIGFM

3.1 任务定义

给定一组多模态样本 $D=(x_1,x_2,\cdots,x_d)$ 作为模型的输入，每个样本 $x_i\in D$ 包含一条由 m 个单词所组成的文本 $S=(w_1,w_2,\cdots,w_m)$、一幅相应的图像 I 以及一个单词数目为 n 的方面实体 $T=(w_1,w_2,\cdots,w_n)$，其中 T 为 S 的单词子序列。本文的研究任务是对每个方面实体在其样本中的情感标签 $y\in Y$ 进行预测，其中 Y 包含积极、消极和中立 3 种类别。

3.2 特征提取模块

本模块采用语言和视觉两类预训练模型分别对单模态的文本特征和图像特征进行提取。

3.2.1 文本特征提取

给定一条输入文本，本文将其分为两个部分：方面实体 T 和其所对应的上下文 C，其中上下文是将文本中方面实体的所在位置使用特殊字符“T”填充获得。对于文本编码，本文采用预训练语言模型 RoBERTa[31] 作为文本编码器，其作为 BERT[32] 模型的扩展，在多种自然语言处理任务中有着较好的表现。具体而言，我们遵循 RoBERTa 编码的实现机制，首先使用“〈/s〉”将 C 和 T 拼接构成上下文感知的方面实体输入 T'，然后对 T 和 T' 分别添加两个特殊标记(即在开头处添加“〈s〉”，结尾处添加“〈/s〉”)送入 RoBERTa 编码后获得方面实体表示和上下文感知的方面实体表示。

$$H_T=RoBERTa(T) \tag{1}$$

$$H_T{}'=RoBERTa(T') \tag{2}$$

其中，$H_T\in\mathbb{R}^{d\times n}$ 且 $H_T{}'\in\mathbb{R}^{d\times t}$，$d$ 为每个单词向量的隐藏维度，n 为 T 所包含的单词个数，t 为 T' 所包含的单词个数。

3.2.2 图像特征提取

对于图像编码，本文采用残差网络(ResNet)[33] 作为模型的图像编码器，相较于先前的 VGG 网络[34]，其使用残差连接解决了随着层数增加所带来的梯度消失问题，在图像识别任务中能够提取更深层次的语义信息。具体而言，首先将给定的输入图像 I 调整为 224×224 像素并记为 I'，然后将预训练的 152 层 ResNet 中最后一个卷积层的输出作为图像的视觉表示：

$$H_I=ResNet(I') \tag{3}$$

其中，$H_I\in\mathbb{R}^{2048\times49}$，49 为将 I' 分割成的 7×7 个具有相同大小的视觉块数目，2048 为每个视觉块的向量维度。为了使图像和文本表示处于相同的语义空间以进行后续的跨模态交互，这里使用线性变换将图像表示转换为与文本表示相同的维度：

$$H_V=W_VH_I+b_V \tag{4}$$

其中，$W_V\in\mathbb{R}^{d\times2048}$ 和 $b_V\in\mathbb{R}^d$ 为可学习线性变换参数。

3.3 特征交互模块

在获得上述方面实体和图像的特征表示之后，为了进一步学习方面实体在图像中的特征，本模块通过引入多模态交互层获得方面实体感知的图像表示。

多模态交互层的内部结构如图 3 所示，其核心技术为多头跨模态注意力机制(Multi-Head Cross-Modal Attention Mechanism，MCATT)。具体而言，首先对方面实体和图像做跨模态特征交互，将上下文感知的方面实体表示 $H_T{}'$ 作为查询向量(Query，$\boldsymbol{Q}$)，将图像表示 H_V 作为键向量(Key，$\boldsymbol{K}$)和值向量(Value，$\boldsymbol{V}$)。跨模态注意力机制第 i 个头部的详细计算过程如下：

$$MCATT_i(H_T{}',H_V)=\mathrm{Softmax}\left(\frac{[W_{Qi}H_T{}']^{\mathrm{T}}[W_{Ki}H_V]}{\sqrt{d/m}}\right)[W_{Vi}H_V]^{\mathrm{T}} \tag{5}$$

其中，m 为 MCATT 中头部的数量，$\{W_{Qi}, W_{Ki}, W_{Vi}\} \in \mathbb{R}^{d/m\times d}$ 分别是第 i 个头部的 Query，Key 和 Value 的可学习线性变换参数。然后，将 m 个头部的 MCATT 输出拼接并进行线性变换获得 MCATT 的最终输出。

$$MCATT(H_T', H_V) = W_m[MCATT_1(H_T', H_V), \cdots, MCATT_m(H_T', H_V)]^{\mathrm{T}} + b_m \quad (6)$$

其中，W_m 和 b_m 为 MCATT 的可学习线性变换参数。最后，将 MCATT 输出经过两个层归一化(LN)和一个前馈网络(FFN)进而获得方面实体感知的图像表示。

$$Z_{T\to V} = LN(H_T' + MCATT(H_T', H_V)) \quad (7)$$

$$H_{T\to V} = LN(Z_{T\to V} + FFN(Z_{T\to V})) \quad (8)$$

其中，$H_{T\to V} \in \mathbb{R}^{d\times t}$ 为最终生成的方面实体感知的图像表示。

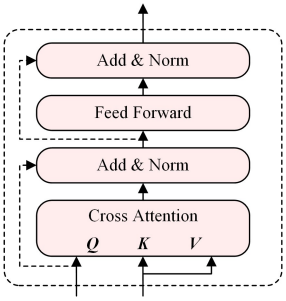


图 3 多模态交互层的内部结构

Fig. 3 Internal structure of multimodal interaction layer

3.4 特征辅助模块

为了使图像中的情感信息能够更直观地进行语义表达，本模块引入 DeepSentiBank[35] 对图像所提取的形容词-名词对(ANPs)从另一个层面生成图像的语义表示。

与上述图像表示不同，ANPs 通过抽取图像中所展现的人或物等名词以及修饰它们的形容词信息，使得模型能够从文本语义层面对图像内容进行理解。具体而言，DeepSentiBank 针对样本中的每幅图像生成一组包含 2 089 个 ANPs 的集合，本文选择置信度排名靠前的 k 个 ANPs 进行实验。

然而，ANPs 的本质是以一种粗粒度方式从图像中提取的内容，其中可能包含与方面实体不相关的图像区域内容或对图像识别错误的语义信息，直接使用这些 ANPs 很可能会因其不准确而为模型引入额外的噪声。在构建多模态方面级情感分析知识增强框架中，Zhao 等[36] 通过计算方面实体和 ANPs 中名词表示之间的语义相似性来获取与方面实体相关的名词，并在实验中取得了良好的对齐效果。本模块受该项工作的启发，首先对上述所选择的 k 个 ANPs 中的形容词和名词分别进行拼接并送入文本编码器获得形容词表示 H_{A} 以及名词表示 H_{N}，然后采用余弦相似度计算方面实体表示 H_{T} 和名词表示 H_{N} 之间的语义相似程度，进而实现方面实体与 ANPs 中名词的对齐。

$$\alpha = \frac{H_{\mathrm{T}}^{\mathrm{T}} \cdot H_{\mathrm{N}}}{\| H_{\mathrm{T}} \| \cdot \| H_{\mathrm{N}} \|} \quad (9)$$

其中，α 为 H_{T} 和 H_{N} 的相似度得分。由于 ANPs 中的形容词在一定程度上有助于反映图像内容所展现的情感，因此将其作为对图像的辅助信息从文本语义层面对方面实体在图像中的情感进行表达。如图 4 所示，本文对 H_{A} 中的每一个形容词个体赋予其所对应的名词和方面实体表示的相似度权重作为图像辅助信息，进而通过文本编码器获得方面实体的图像辅助信息表示。

$$H_{\mathrm{T}\to\mathrm{A}} = \alpha \cdot H_{\mathrm{A}} \quad (10)$$

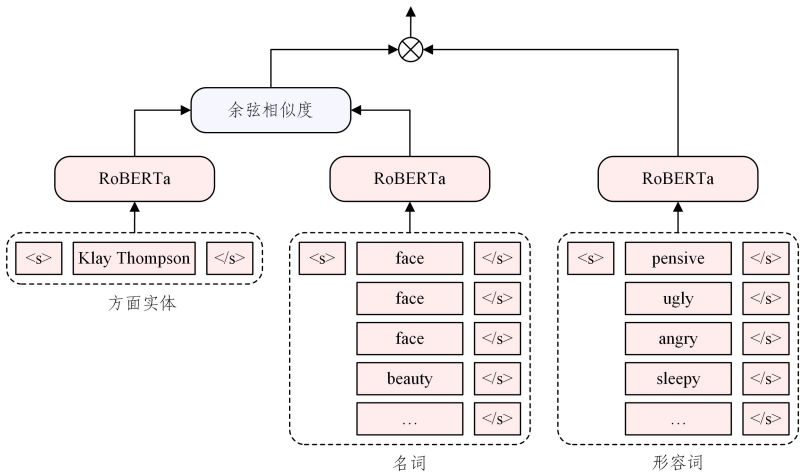


图 4 特征辅助模块的内部结构

Fig. 4 Internal structure of feature auxiliary module

3.5 特征融合模块

本模块将上述获得的上下文感知的方面实体表示 H_{T}'、方面实体感知的图像表示 $H_{\mathrm{T}\to\mathrm{V}}$，以及方面实体的图像辅助信息表示 $H_{\mathrm{T}\to\mathrm{A}}$ 通过指定的方式融合，获得方面实体的最终表示，进而实现情感预测。

本文将图像用作辅助文本语义表达的另一种模态信息，其在一定程度上虽然能够为模型提供文本中所不包含的一些内容，但其中可能存在与文本语义不相关的区域，进而为模型引入额外的噪声。此外，图像辅助信息中也可能存在形容词识别错误等情况，因此本文在多模态特征融合阶段构建一种针对文本以外信息的门控机制进而动态控制这些特征表示的输入。具体而言，首先将 H_{T}' 和 $H_{\mathrm{T}\to\mathrm{V}}$ 进行拼接，通过线性变换和非线性激活函数构建动态控制图像信息输入的门控机制：

$$g_{T\to V}=\sigma(W_{T\to V}[H_T';H_{T\to V}]+b_{T\to V}) \tag{11}$$

其中，$W_{T\to V}\in\mathbb{R}^{d\times 2d}$ 和 $b_{T\to V}\in\mathbb{R}^{d}$ 为可学习线性变换参数；σ 为使用点乘的非线性激活函数，其用于将门控的输出控制在[0,1]的范围内。同理，使用上述构建方法也能够获得动态控制图像辅助信息输入的门控机制。

$$g_{T\to A}=\sigma(W_{T\to A}[H_T';H_{T\to A}]+b_{T\to A}) \tag{12}$$

其中，$W_{T\to A}\in\mathbb{R}^{d\times 2d}$ 和 $b_{T\to A}\in\mathbb{R}^{d}$ 为可学习线性变换参数。接着，将图像以及图像辅助信息的门控生成值分别与方面实体感知的图像表示以及方面实体的图像辅助信息表示进行对应元素相乘，实现以文本中的单词级力度动态控制图像和图像辅助信息的输入。然后，将上述表示进行拼接并送入基于 Transformer 的多模态自注意力层实现模态间的特征融合。

$$H=Transformer(H_T';g_{T\to V}H_{T\to V};g_{T\to A}H_{T\to A}) \tag{13}$$

最后，将多模态特征融合表示 H 的第一个 token 表示 H^0 送入 Softmax 层，进而获得最终的情感标签。

$$P(y\mid H)=\mathrm{Softmax}(WH^0+b)$$

为了对模型中的参数进行优化，本文采用方面实体情感标签的预测值与其真实值所构建的交叉熵损失作为模型情感分析任务的训练损失函数。

$$L=-\frac{1}{|D|}\sum_{j=1}^{|D|}\log P(y^j\mid H^0) \tag{14}$$

4 实验

为了证明 TIGFM 模型的有效性，本文在两个基于多模态方面级情感分析的 Twitter 数据集上开展了一系列的实验，并选择一些具有代表性的方法与 TIGFM 模型进行性能上的对比。

4.1 实验设置

本文选用由 Yu 等[5]提出的两个基于多模态方面级情感分析的数据集 Twitter-2015 和 Twitter-2017 进行实验，这两个数据集采样自 Twitter 平台在 2014－2015 年和 2016－2017 年所发布的包含文本和图像的推文，相关统计信息如表 1 所列(其中 Pos 代表积极，Neg 代表消极，Neu 代表中立)。

表 1 两个 Twitter 数据集的相关统计信息

Table 1 Statistics of two Twitter datasets

	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Pos	928	303	317	1508	515	493
Neg	368	149	113	416	144	168
Neu	1883	670	607	1638	517	573
Total	3179	1122	1037	3562	1176	1234

实验中采用预训练的 RoBERTa-base[31]模型作为文本编码器，ResNet-152[33]作为图像编码器。此外，在交替优化过程中采用 AdamW 作为学习器对参数进行优化。对于 TIGFM 模型的超参数，本文将 batch size 值设置为 16，训练 epoch 值设置为 9，k 值设置为 5，学习率设置为 1×10^{-5}。本文选用对 TIGFM 模型 3 次独立训练的平均结果作为最终的实验结果，所有实验均基于 PyTorch 以及 NVIDIA TeslaV100 显卡实现。

4.2 对比基线

本文选用表 2 中的单模态和多模态模型作为基线与 TIGFM 模型进行性能上的对比。

表 2 对比基线

Table 2 Comparison baselines

模型	简介
Res-Target	直接通过 ResNet 模型获得图像视觉特征
AE-LSTM[37]	通过注意力获得方面实体相关上下文的 LSTM 模型
MGAN[38]	以不同粒度融合方面实体和上下文的多粒度注意力网络模型
BERT[32]	基于 Transformer 用于方面实体和文本之间交互的预训练语言模型
RoBERTa[31]	采用更好的训练策略和更大的语料库对 BERT 进行进一步改进的预训练语言模型
MIMN[3]	用于方面实体、文本和图像信息之间交互的多跳记忆网络模型
ESAFN[4]	用于捕获方面实体、文本和图像信息之间关系的实体感知注意力融合网络模型
TomBERT[5]	基于 BERT 改进的多模态方面级情感分析模型
CapBERT[30]	将图像转换为文本字幕并与输入文本相结合进行编码的多模态方面级情感分析模型
CapRoBERTa	使用 RoBERTa 代替 CapBERT 中 BERT 的扩展基线

4.3 实验结果与分析

表 3 列出了 TIGFM 模型和对比基线在 Twitter-2015 和 Twitter-2017 数据集上的实验结果，本文采用准确率(Accuracy，Acc)和 F1 值(Macro-F1，F1)作为模型性能的评价指标。TIGFM 模型在两个数据集上均取得了最优结果，相较于基线中性能最佳的模型，在 Twitter-2015 数据集上的 Acc 和 F1 分别提高约 0.7%和 0.5%，在 Twitter-2017 数据集上的 Acc 和 F1 分别提高约 1.1%和 2.0%。

表 3 TIGFM 模型和对比基线在两个 Twitter 数据集上的实验结果

Table 3 Experimental results of TIGFM and comparison baselines on two Twitter datasets

(%)

模态	模型	Twitter-2015		Twitter-2017	
		Acc	*F1*	*Acc*	*F1*
图像	Res-Target	59.88	46.48	58.59	53.98
文本	AE-LSTM	70.30	63.43	61.67	57.97
	MGAN	71.17	64.21	64.75	61.46
	BERT	74.15	68.86	68.15	65.23
	RoBERTa	76.28	71.36	69.77	68.00
图像＋文本	MIMN	71.84	65.69	65.88	62.99
	ESAFN	73.38	67.37	67.83	64.22
	TomBERT	77.15	71.75	70.34	68.03
	CapBERT	78.01	73.25	69.77	68.42
	CapRoBERTa	77.82	73.38	71.07	68.57
	TIGFM	**78.66**	**73.89**	**72.12**	**70.58**

针对表 3 中的实验结果，可以得出以下结论：(1)Res-Target 模型的性能低于所有其他模型的性能，这表明大多数图像都可以对文本起到辅助作用，而无法作为独立模态主导模型的情感预测；(2)融合图像信息的多模态模型整体上比单模态模型的性能更好，这表明图像信息能够对文本加以补充，从而增强模型的情感预测能力；(3)TomBERT 和 CapBERT 模型性能要远优于其他多模态模型的性能，这表明使用 BERT 作为文本编码器可以获得更具鲁棒性的特征表示；(4)CapBERT 模型在所有原始基线中取得了最佳的性能，这表明采用图像字幕的文本信息相较于图像有着更为直观的

语义表示；(5)CapRoBERTa 模型在以上评价指标中的整体性能优于 CapBERT 模型的性能，这也证明了 RoBERTa 比 BERT 的性能更为强大。

4.4 消融实验

为了进一步研究 TIGFM 模型中的各个单元对其整体性能的影响，本节在 Twitter-2015 和 Twitter-2017 数据集上对模型中的 3 个重要单元进行消融分析：(1)图像辅助信息(Image Auxiliary Information，IAI)；(2)图像门控机制(Image Gating Mechanism，IGM)；(3)图像辅助信息门控机制(Image Auxiliary Information Gating Mechanism，IAIGM)。实验结果如表 4 所列，首先分别移除上述 3 个单元，然后再将这 3 个单元全部移除，进而更为全面地验证本文所设计的单元对模型性能提升所做出的贡献。

表 4 TIGFM 模型的消融实验结果

Fig. 4 Ablation results of TIGFM (%)

单元	Twitter-2015		Twitter-2017	
	Acc	*F1*	*Acc*	*F1*
TIGFM	**78.66**	**73.89**	**72.12**	**70.58**
移除 IAI	77.37	72.32	71.61	69.80
移除 IGM	78.24	73.26	71.96	70.39
移除 IAIGM	78.21	73.21	71.74	70.18
移除以上全部	77.18	71.90	71.56	69.02

移除 IAI 后模型在两个数据集上的 Acc 分别下降约 1.3%和 0.5%，这表明使用 ANPs 中形容词的加权作为图像辅助信息可以更为直观地对图像中的情感信息进行语义表达；移除 IGM 以及 IAIGM 后模型在两个数据集上的 Acc 分别下降约 0.4%，0.2%以及 0.5%，0.4%，这表明图像和 ANPs 中的形容词虽然在一定程度上能够提供文本以外的内容，但是直接使用它们也会因其中存在与文本语义不相关的信息而为模型引入额外的噪声；最后，移除上述全部单元后模型在两个数据集上的 Acc 分别下降约 1.5%和 0.6%，这从另一个角度验证了本文所设计的单元对模型性能提升均有一定程度的贡献。

4.5 参数设置

上述实验均为基于调优后的模型超参数所设置的，本节对评估最优超参数的过程进行详细分析。

4.5.1 batch size 取值

表 5 列出了在两个数据集上不同 batch size 的取值对模型性能的影响，实验分别使用 8，16 和 32 这 3 个数值作为 batch size 的取值进行分析。

表 5 batch size 取值对模型性能的影响

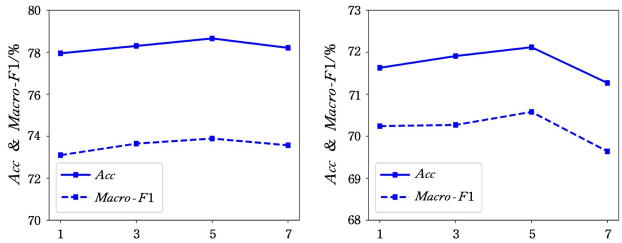
Table 5 Effect of batch size on model performance (%)

batch size 取值	Twitter-2015		Twitter-2017	
	Acc	*F1*	*Acc*	*F1*
8	77.47	73.16	71.82	70.44
16	**78.66**	**73.89**	**72.12**	**70.58**
32	78.08	73.25	71.94	70.51

从表 5 中的结果可以看出，当 batch size 为 16 时，模型在两个数据集上均取得了最佳性能，我们推测原因如下：当 batch size 为 8 时，对于两个数据集中的样本数量而言该 batch size 取值偏小，模型不但训练耗时较长而且难以收敛，进而导致欠拟合；在一定范围内提高 batch size 有利于模型收敛的稳定性，但是当 batch size 为 32 时，模型可能因该 batch size 取值过大而陷入局部极小值进而导致泛化性能下降。因此，最终将 batch size 的取值设置为 16。

4.5.2 *k* 取值

为了探索 ANPs 数量 *k* 的最优取值，实验分别使用 1，3，5 和 7 这 4 个数值作为 *k* 的取值进行分析，图 5(a)和图 5(b)分别展示了在两个数据集上 *k* 的取值对模型性能的影响。



(a)Twitter-2015 (b)Twitter-2017

图 5 *k* 取值对模型性能的影响

Fig. 5 Effect of *k* on model performance

从图 5 中的结果可以看出，随着 ANPs 数量的增加，模型性能呈现上升的趋势，当 *k* 为 5 时模型在两个数据集上均取得了最佳性能，然而当 *k* 大于 5 时模型性能不再提升反而下降，我们推测原因如下：数据集中的每条文本所涉及的方面实体可能最多不超过 5 个，当 *k* 大于文本中方面实体的数量时，图像辅助信息可能会为模型引入额外的噪声。因此，最终将 *k* 的取值设置为 5。

4.6 样例分析

为了更好地展示本文所提出的模型在两个数据集上的性能优势，本节选用两个具有代表性的样例将 TIGFM 与 CapRoBERTa 模型进行对比分析，样例信息和预测结果如表 6 所列。

表 6 CapRoBERTa 和 TIGFM 模型的样例分析

Table 6 Case analysis of CapRoBERTa and TIGFM

图像		
文本	(a) RT @ BBCOne: Dear [Madonna]$_{Pos}$, THIS is how you wear a cape. # Poldark # Demelza	(b) pitchfork: Tune into ArianaGrande ' s "[One Love Manchester]$_{Pos}$" benefit concert live stream, which has begun …
ANPs	young muslim bright stars christian saint young adult holy cross	sexy dance christian concert empty trash energetic performance favorite club
真实预测	(Madonna, Pos)	(One Love Manchester, Pos)
CapRoBERTa	(Madonna, Neu)	(One Love Manchester, Neu)
TIGFM	(Madonna, Pos)	(One Love Manchester, Pos)

表 6 样例(a)中，TIGFM 模型结合方面实体“Madonna”在图像中微笑的面部表情和 ANPs 中“bright”等具有积极色彩的形容词预测出方面实体积极的情感极性；表 6 样例(b)中，由于方面实体“One Love Manchester”在图像中的表示为整个场景，结合场景中的绚丽舞台和 ANPs 中“energetic”等

同样具有积极色彩的形容词，TIGFM 模型也能够预测出方面实体积极的情感极性。CapRoBERTa 模型在两个样例中均做出了错误的预测。通过调研得知样例(a)和(b)的图像字幕信息“A group of people in traditional dress clothes”和“A woman in a blue dress is talking on a cell phone”均不存在具有情感色彩的词汇，并且 CapRoBERTa 模型在获得图像语义的过程中仅依赖于这些字幕信息而丢弃了原始图像本身，因此其在一定程度上并不能准确反映图像中所体现的情感。

结束语 本文提出了一种基于文本和图像门控融合机制的多模态方面级情感分析模型(TIGFM)。该模型通过引入从数据集图像中提取的形容词-名词对(ANPs)，并将其中形容词的加权作为图像辅助信息，使得样本中的图像内容能够获得更好的情感语义表达，此外在最后的特征融合阶段，通过构建一种动态控制图像信息输入的多模态特征融合门控机制，进而防止与文本语义不相关的信息为模型引入额外的噪声。实验结果表明，在 Twitter-2015 和 Twitter-2017 数据集上 TIGFM 模型的性能优于所有对比基线模型，进而验证了本文所提出模型的优越性和方法的有效性。然而，本文研究仍存在以下问题：(1)该模型仅在多模态特征融合阶段设计了一种门控机制，在此之前并未对原始文本和图像的语义相关性做出判断；(2)实验所采用数据集的数量较少且类型较为单一，对于除英文之外的数据集模型情感预测性能尚未知晓。后续的工作将着重对上述两项问题展开研究。

未来，我们计划将 TIGFM 模型应用到更多与多模态相关的研究任务中以证明本文所提方法的有效性和普适性，同时也将进一步构建中文社交媒体或服务平台数据集以确保模型在多种语言或场景任务上的适用性。此外，随着近期大模型热度的不断提升，在接下来的研究任务中同样计划探索如何有效结合大模型至现有的工作中，并进一步实现更为具体的情感多分类任务。

参 考 文 献

[1] CHEEMA G S, HAKIMOV S, MÜLLER-BUDACK E, et al. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods[C]// Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding. 2021: 37-45.

[2] YUAN J L, DING Y Y, SHENG D M, et al. Image-Text Sentiment Analysis Model Based on Visual Aspect Attention[J]. Computer Science, 2022, 49(1): 219-224.

[3] XU N, MAO W, CHEN G. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 371-378.

[4] YU J, JIANG J, XIA R. Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 429-439.

[5] YU J, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019: 5408-5414.

[6] WANG J, LIU Z, SHENG V, et al. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification[C]// Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29, November 1, 2021, Proceedings, Part III 4. Springer International Publishing, 2021: 3-15.

[7] 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.

[8] PORIA S, HAZARIKA D, MAJUMDER N, et al. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research[J]. IEEE Transactions on Affective Computing, 2020, 14: 108-132.

[9] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers). 2017: 873-883.

[10] LIANG P P, LIU Z, ZADEH A, et al. Multimodal Language Analysis with Recurrent Multistage Fusion[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 150-161.

[11] BUSSO C, DENG Z, YILDIRIM S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information[C]// Proceedings of the 6th International Conference on Multimodal Interfaces. 2004: 205-211.

[12] LEE C C, MOWER E, BUSSO C, et al. Emotion recognition using a hierarchical binary decision tree approach[J]. Speech Communication, 2011, 53(9/10): 1162-1171.

[13] CASTRO S, HAZARIKA D, PÉREZ-ROSAS V, et al. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper) [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4619-4629.

[14] CAI Y, CAI H, WAN X. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2506-2515.

[15] YANG J, SHE D, SUN M, et al. Visual sentiment prediction based on automatic discovery of affective regions[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2513-2525.

[16] YANG J, SHE D, LAI Y K, et al. Weakly supervised coupled networks for visual sentiment analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7584-7592.

[17] KUMAR A, GARG G. Sentiment analysis of multimodal twitter data[J]. Multimedia Tools and Applications, 2019, 78: 24103-24119.

[18] KUMAR A, SRINIVASAN K, CHENG W H, et al. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data[J]. Information Processing & Management, 2020, 57(1): 102141.

[19] 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.

[20] PONTIKI M,GALANIS D,PAPAGEORGIOU H,et al. Semeval-2016 task 5: Aspect based sentiment analysis [C] // ProWorkshop on Semantic Evaluation(SemEval-2016). Association for Computational Linguistics,2016:19-30.

[21] DONG L,WEI F,TAN C,et al. Adaptive recursive neural network for target-dependent twitter sentiment classification[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers). 2014:49-54.

[22] XUE W,LI T. Aspect Based Sentiment Analysis with Gated Convolutional Networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers). 2018:2514-2523.

[23] MA Y,PENG H,CAMBRIA E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018:5876-5883.

[24] MEŠKELČ D,FRASINCAR F. ALDONAr:A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model[J]. Information Processing & Management,2020,57(3):102211.

[25] ZHAO L,LIU Y,ZHANG M,et al. Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model[J]. Information Processing & Management,2021,58(5):102641.

[26] WANG K,SHEN W,YANG Y,et al. Relational Graph Attention Network for Aspect-based Sentiment Analysis[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:3229-3238.

[27] ZHANG M,QIAN T. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis[C]// Proceedings of the 2020 Conference on Empirical Methods in NaturalLanguage Processing(EMNLP). 2020:3540-3549.

[28] XU H,LIU B,SHU L,et al. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Volume 1(Long and Short Papers). 2019:2324-2335.

[29] SUN C,HUANG L,QIU X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence[C]// Proceedings of NAACL-HLT. 2019:380-385.

[30] KHAN Z,FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:3034-3042.

[31] LIU Y,OTT M,GOYAL N,et al. Roberta:A robustly optimized bert pretraining approach[J]. arXiv:1907.11692,2019.

[32] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT. 2019:4171-4186.

[33] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

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

[35] CHEN T,BORTH D,DARRELL T,et al. Deepsentibank:Visual sentiment concept classification with deep convolutional neural networks[J]. arXiv:1410.8586,2014.

[36] ZHAO F,WU Z,LONG S,et al. Learning from Adjective-Noun Pairs:A Knowledge-enhanced Framework for Target-Oriented Multimodal Sentiment Classification [C] // Proceedings of the 29th International Conference on Computational Linguistics. 2022:6784-6794.

[37] WANG Y,HUANG M,ZHU X,et al. Attention-based LSTM for aspect-level sentiment classification[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:606-615.

[38] 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.

ZHANG Tianzhi, born in 1995, postgraduate. His main research interests include sentiment analysis and data mining.

ZHOU Gang, born in 1974, Ph.D, professor. His main research interests include mass data processing and knowledge grapgh.

(责任编辑:何杨)