基于可信细粒度对齐的多模态方面级情感分析

范东旭, 过弋

引用本文

范东旭, 过弋. 基于可信细粒度对齐的多模态方面级情感分析[J]. 计算机科学, 2023, 50(12): 246-254.

FAN Dongxu, GUO Yi. Aspect-based Multimodal Sentiment Analysis Based on Trusted Fine-grained Alignment [J]. Computer Science, 2023, 50(12): 246-254.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100038

基于可信细粒度对齐的多模态方面级情感分析

范东旭[1]　过　弋[1,2,3]
1 华东理工大学信息科学与技术学院　上海 200237
2 大数据流通与交易技术国家工程实验室-商业智能与可视化技术研究中心　上海 200436
3 上海大数据与互联网受众工程技术研究中心　上海 200072
(956701698@qq.com)

摘　要　基于方面的多模态情感分析任务(Multimodal Aspect-Based Sentiment Analysis,MABSA),旨在根据文本和图像信息识别出文本中某特定方面词的情感极性。然而,目前主流的模型并没有充分利用不同模态之间的细粒度语义对齐,而是采用整个图像的视觉特征与文本中的每一个单词进行信息融合,忽略了图像视觉区域和方面词之间的强对应关系,这将导致图片中的噪声信息也被融合进最终的多模态表征中,因此提出了一个可信细粒度对齐模型 TFGA(MABSA Based on Trusted Fine-grained Alignment)。具体来说,使用 FasterRCNN 捕获到图像中包含的视觉目标后,分别计算其与方面词之间的相关性,为了避免视觉区域与方面词的局部语义相似性在图像文本的全局角度不一致的情况,使用置信度对局部语义相似性进行加权约束,过滤掉不可靠的匹配对,使得模型重点关注图片中与方面词相关性最高且最可信的视觉局域信息,降低图片中多余噪声信息的影响;接着提出细粒度特征融合机制,将聚焦到的视觉信息与文本信息进行充分融合,以得到最终的情感分类结果。在 Twitter 数据集上进行实验,结果表明,文本与视觉的细粒度对齐对方面级情感分析是有利的。

关键词:方面级情感分析;多模态;细粒度对齐;情感分析;自然语言处理
中图法分类号　TP391.1;TP18

Aspect-based Multimodal Sentiment Analysis Based on Trusted Fine-grained Alignment

FAN Dongxu[1] and GUO Yi[1,2,3]
1 School of Computer Science and Engineering,East China University of Science and Technology,Shanghai 200237,China
2 Business Intelligence and Visualization Research Center,National Engineering Laboratory for Big Data Distribution and Exchange Technologies,Shanghai 200436,China
3 Shanghai Engineering Research Center of Big Data & Internet Audience,Shanghai 200072,China

Abstract　Aspect based multimodal sentiment analysis task(MABSA) aims to identify the sentiment polarity of a specific aspect word in a text based on text and image information. However,the current mainstream model does not make full use of the fine-grained semantic alignment between different modes. Instead,it uses the image features of the entire image to fuse information with each word in the text,ignoring the strong correspondence between the local image information and aspect words,which will lead to the noise information in the image being integrated into the final multimodal representation,Therefore,this paper proposes a trusted fine-grained alignment model TFGA(MABSA based on trusted fine-grained alignment). Specifically,we use FasterRCNN to capture the visual objects contained in the image,and then calculate the correlation between them and aspect words respectively. To avoid the inconsistency of the local semantic similarity between the visual object and aspect words in the global perspective of the image-text,confidence is used to weight the local semantic similarity and filter out the unreliable matching pairs,then the model can focuse on the most reliable and highest visual local information related to aspect words in the image to reduce the impact of redundant noise information in the image. Then a fine-grained feature fusion mechanism is proposed to fully fuse the focused local image information with the text information to obtain the final sentiment classification result. Experiments on Twitter datasets show that fine-grained alignment of text and vision is beneficial to aspect based sentiment analysis.

Keywords　Aspect-based sentiment analysis,Multimodal,Fine-grained alignment,Sentiment analysis,Natural language processing

到稿日期:2022-11-06　返修日期:2023-03-09
基金项目:上海市科学技术委员会科技计划项目(22DZ204903,22511104800)
This work was supported by the Science and Technology Plan Project of Shanghai Municipal Commission of Science and Technology (22DZ204903,22511104800).
通信作者:过弋(guoyi@ecust.edu.cn)

1 引言

随着社交媒体的发展，互联网中产生了大量的包含用户丰富的情感和意见的信息，这些信息以多种格式表示，例如评论、标签等。从这些信息中挖掘情感信息，对情感分析、意见挖掘领域发挥着至关重要的作用，它可以预测人类的决策，并支持一些实际应用，如民意分析、品牌监测和政治投票预测等。然而，到目前为止，基于方面的情感分析仍然更加关注纯文本信息，而社交媒体用户越来越多地使用附加图像来表达自己的体验和观点，图像正成为网络中的一种关键的数据类型，因此从丰富的文本和视觉内容中提取用户对不同方面的情感至关重要。

多模态方面级情感分类（MABSA）作为细粒度情感分析任务，相对于篇章级或句子级的情感预测来说，其更偏向于从用户评价中分析出用户对产品某个属性或者事件的某个元素的情感。如图 1 左侧所示，预计将从文本中提取出方面词 Takis Fuego 的情感极性为 Positive。近两年虽然有一些基于深度学习的方法在 MABSA 任务上取得了不错的效果，但为了进一步提升模型的性能，一些研究为图像信息设置了门控机制，从而判断图像是否提供有用信息。Yu 等[1]提出了 ES-AFN 模型，通过门控机制从整体角度来消除视觉噪声，并取得了不错的成果，这初步表明将图像信息进行筛选对分类效果的提高有一定的帮助。但对于完整的图像特征，设置门控机制的方式难免会将部分有用的信息筛选掉，因此如何实现细粒度的图像过滤是细粒度情感分析任务的一个重点。

细粒度对齐对于 MABSA 任务至关重要，因为直接将完整的图片特征向量和文本特征向量进行交互计算，忽略了文本中的方面词只与图片中的部分信息有较强的对应关系。一般情况下，图像中包含较多的视觉目标（细粒度图像表示），通过视觉目标与方面词的对齐来捕获图像的有用部分，从而减轻视觉噪声，有利于情感分类性能的提升。例如图 1 右侧中，绿色边界框内的视觉信息为方面词 Lady Gaga 的情感分类提供了最重要的信息，但却有可能会导致其余的两个方面词也被错误地预测为 positive。因此如何通过图像和方面词细粒度对齐的方式，达到在过滤噪声的同时还能捕获到对情感分类有用的局部信息的目的，是现阶段研究的一个难点。当图像中包含较多与方面词信息相关的视觉目标时，例如图 1 右侧中绿色和黄色的 3 个视觉框内都包含 people 类视觉目标，如何判断出哪一个是真正的 Lady Gaga，完成更准确可信的语义对齐是该任务的另一个难点。

图 1 多模态方面级情感分析的例子（电子版为彩图）

Fig. 1 Example of multimodal aspect-based sentiment analysis

针对以上问题，本文在以往研究的基础上提出了一种新的基于可信细粒度对齐的多模态方面级情感分析模型（TFGA）。通过计算细粒度图像与方面词之间的相关性，来过滤视觉噪声信息。然而，直接使用视觉区域和方面词的局部语义相似性进行过滤，从图像和文本的全局角度来看，它可能是不可靠的，从而导致不准确的相关性测量。因此，为了解释视觉区域与方面词的局部语义相似性对整体图文关联度的真实贡献水平，本文引入相似性匹配置信度来约束相关性的计算，该置信度通过整体图文语义相似性与视觉区域和方面词的相似性之间的内积来衡量。

本文的主要贡献如下：

1）针对图像中噪声信息会限制模型性能的问题，使用 FasterRCNN[2]模型从图片中抽取出不同视觉实体，采用基于距离的相关性计算方法，并通过视觉区域与方面词的局部语义相似性对整体图文关联度的真实贡献水平来约束相关性的计算，使得模型聚焦到可信的、有用的局部图像区域，从而过滤与文本和方面词无关的噪声图像信息。

2）考虑到文本和视觉模态之间信息的协同互补，本文使用了细粒度双线性融合机制，使方面词、文本、图像、视觉区域四者之间充分交互融合，来提取用于情感分类的重要交互特征。

3）本文提出的 TFGA 模型在 MABSA 任务的两个公开基准数据集 TWITTER-2015 和 TWITTER-2017 上进行实验，并与相关模型进行对比，结果表明，本文模型的性能优于几种高度竞争的单模态和多模态方法，证明了本文提出的可信细粒度对齐机制的有效性。

2 相关工作

2.1 文本方面级情感分析

传统的方面级情感分析模型主要是针对文本信息，已经在 NLP 领域进行了广泛研究。早期的方法主要集中于机器学习[3-4]，该方法的性能依赖于特征标注的好坏，因此需要投入大量的人力物力。随着深度学习技术的进步，这一问题逐渐得到改善，并带来了性能的大幅提升。受到句子的不同部分对特定方面发挥不同作用的启发，文献[5-9]设计了不同的注意力机制来抽取方面词的表征。Ma 等[10]提出了采用更复杂的交互注意力机制对方面词和文本进行协同建模的 IAN 模型。Fan 等[11]提出了一个捕获多层次方面词文本交互的多粒度注意力网络。除了循环神经网络，还探索了其他网络结构与注意力机制的结合，如 CNN 网络[12-13]、门控网络[12,14]、记忆网络。近几年，预训练语言模型逐渐成为主流[15-17]。虽然基于纯文本的方面级情感分析取得了很大的成功，但它们只依赖文本信息，似乎不能有效地处理信息多样的社交媒体数据，因此多模态方面级情感分析逐渐走入科研人员的视野。

2.2 文本＋视觉方面级情感分析

由于社交媒体的日益普及，多模态方面级情感分析任务受到了越来越多的关注。针对该任务，近几年涌现出了一些深度学习方法。同样受到注意力机制[18]在其他自然语言处理任务中获得上下文信息优势的启发，Yu 等[1]、Xu 等[19]和

Liu 等[20]设计了不同的有效注意力机制来建模方面词、文本和图像之间的交互。Yu 等[21]设计了名为 TwitterBERT 的模型，结合预训练和微调，调整了现有的预训练语言模型 bert 来捕获文本和图像之间的交互，获得了较为出色的效果。Yu 等[22]提出了基于多模态提示的微调方法，用于解决不同粒度的情感预测任务。Zhao 等[23]通过从图像中提取形容词-名词对来帮助模型进行文本和图像对齐。Fu 等[24]提出了基于 Transformer 的模型，将图像翻译为辅助句子，将原始句子和辅助句子相结合来进行有针对性的情感分类的方法。Yu 等[25]设计了分层交互式多模态变压器来捕获文本和图像的交互信息并消除两者间的语义差异。Ju 等[26]提出用端到端的方式联合提取方面词及其情感极性。这些研究有以下的两个特点：(1)忽略了方面词和图像局部信息的强对应关系；(2)图像和文本整体特征进行融合，没有进行充分的信息交互。基于此提出本文模型，通过实现图文之间可信细粒度对齐来过滤噪声信息，并进行图文细粒度融合交互，提高了情感分类的性能。

3 TFGA 模型

本文模型主要分为 4 个部分，即特征抽取、图文细粒度对齐、模态融合和情感标签分类，模型的整体结构如图 2 所示。对于给定的多模态训练样本数据集 D，其中每一个样本 $d\in D$，包含一个有 n 个单词的句子 $T=\{w_1,w_2,w_3,\cdots,w_n\}$ 和一个与该句子相关的图片 I，以及一个包含 m 个单词的方面项 $A=\{w_{a+1},w_{a+2},w_{a+3},\cdots,w_{a+m}\}$，方面项是句子 L 的一部分，其中 a 为方面项起始单词在文本中的位置。样本中的每一个方面项，都有一个对应的情感极性标签 y，$y\in\{positive,negative,neutral\}$。本文模型的任务是将 D 作为训练数据集，训练一个模型可以根据 T 和 I 准确地判断出样本中方面项 A 的情感极性。

3.1 特征抽取

3.1.1 文本表示

使用一个预训练好的词嵌入矩阵 ***Glove*** 来获得每一个单词固定的初始词嵌入向量，假如词嵌入矩阵为 $\boldsymbol{M}\in\boldsymbol{R}^{d*|V|}$，其中 d 是词向量的维度，$|V|$ 是词典大小，文本中每一个单词对应 $\boldsymbol{M}$ 矩阵中的一行。转换后的句子表示为 $\widetilde{\boldsymbol{H}}=\{\widetilde{\boldsymbol{H}}_1,\widetilde{\boldsymbol{H}}_2,\cdots,\widetilde{\boldsymbol{H}}_n\}$。其中 $\widetilde{\boldsymbol{H}}_i\in\boldsymbol{R}^{|V|}$，将 $\widetilde{\boldsymbol{H}}$ 送入双向 LSTM 来获得文本的上下文依赖关系，并把其最后一层的隐藏状态作为最终文本向量表示 $\boldsymbol{H}=\{\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_n\}$。如果方面词由多个单词组成，则取所有单词的词嵌入的平均值作为最终方面词的向量表示 $\boldsymbol{H}_{avg}$。基于这些隐藏状态，进一步采用广泛使用的注意力机制[18]来计算文本的全局表示，使用 $\boldsymbol{H}_{avg}$ 作为 query，其计算过程如下：

$$\widetilde{\boldsymbol{H}}_k=\boldsymbol{M}(w_k),k\in[1,n] \tag{1}$$

$$\overleftarrow{\boldsymbol{H}}_k=\overleftarrow{LSTM}(\widetilde{\boldsymbol{H}}_k,\overleftarrow{\boldsymbol{H}}_{k-1}),k\in[1,n] \tag{2}$$

$$\overrightarrow{\boldsymbol{H}}_k=\overrightarrow{LSTM}(\widetilde{\boldsymbol{H}}_k,\overrightarrow{\boldsymbol{H}}_{k-1}),k\in[1,n] \tag{3}$$

$$\boldsymbol{H}_k=concat(\overleftarrow{\boldsymbol{H}}_k,\overrightarrow{\boldsymbol{H}}_k),k\in[1,n] \tag{4}$$

$$\boldsymbol{H}_{avg}=\sum_{i=1}^{m}\boldsymbol{H}_{a+i}/m \tag{5}$$

$$\boldsymbol{T}^{glo}=\sum_{i=1}^{n}w_i\boldsymbol{H}_i/\|\sum_{i=1}^{n}w_i\boldsymbol{H}_i\|_2 \tag{6}$$

其中，$\overrightarrow{\boldsymbol{H}}_k$，$\overleftarrow{\boldsymbol{H}}_k$ 分别表示双向 LSTM 第 k 层正向与反向隐藏状态，注意力权重 w_i 是 $\boldsymbol{H}_i$ 和 $\boldsymbol{H}_{avg}$ 之间的归一化相似性。

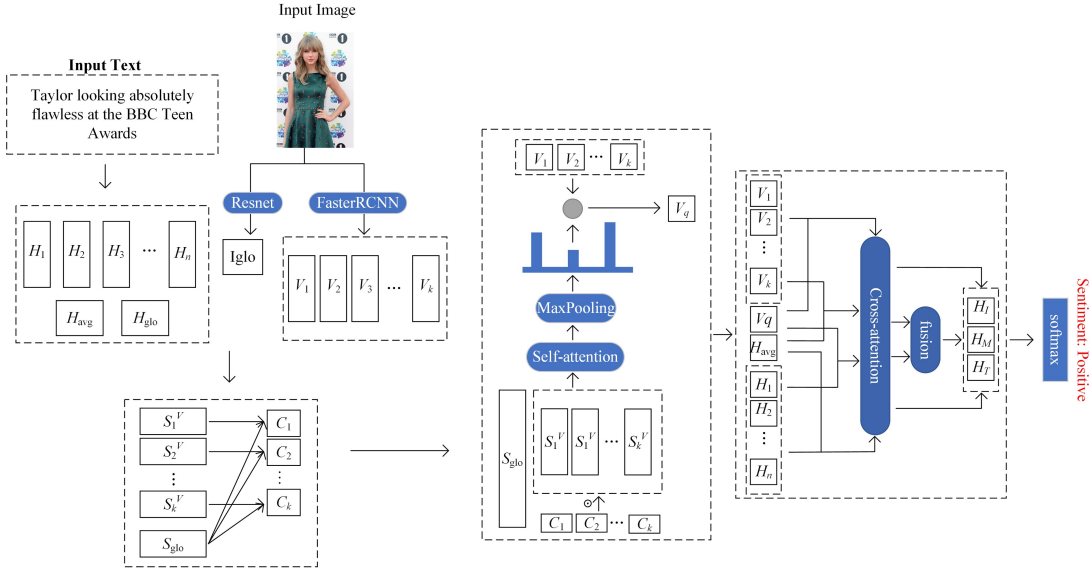


图 2 TFGA 模型的整体结构

Fig. 2 Overall structure of TFGA model

3.1.2 视觉表示

由于深度 CNN 模型在大部分图像处理任务中具有良好的性能，能够捕获对任务有用的高级特征，因此本文采取了一种较为先进的 CNN 模型，即剩余网络(ResNet)，用于提取图片的视觉特征。对于输入图片 I，首先将其大小调整为固定 224 * 224 格式，以适应网络的输入需求，然后将转换后的图片送入 ResNet 模型中，使用模型最后一个卷积层的输出作为图片视觉特征表示 $\widetilde{\boldsymbol{R}}$，最后使用线性转换函数 $\boldsymbol{R}=\boldsymbol{W}_I\widetilde{\boldsymbol{R}}$ 将视觉特征投影到与文本特征的相同空间中，其中 $\boldsymbol{W}_I\in\boldsymbol{R}^{d*2048}$，$\boldsymbol{R}\in\boldsymbol{R}^{d*49}$，计算过程如下：

$$\widetilde{\boldsymbol{R}}=\mathrm{Res}Net(I) \tag{7}$$

$$ResNet(I)=\{r_i \mid r_i\in R^{2048}, j=1,2,\cdots,49\} \tag{8}$$

可以观察到 $\widetilde{\boldsymbol{R}}\in\boldsymbol{R}^{49*2048}$，其中 49 为图片区域的数量，然而方面词与图像中的对象有很强的一致性，而与其他区域没有关系，在所有的区域使用注意力机制不仅会引入噪声，而且会导致模型更难从图像中提取出有用的特征。因此，为了从图片中提取出对象级图片信息，本文使用预先训练好的 Faste-RCNN[2] 目标检测模型来检测图像中的显著区域。通常情况下，只有图像中的较为显著的区域与文本信息相关，因此只取分类得分较高的前 k 个图像区域，准确地说是非最大抑制处理后的前 k 个视觉实体区域$\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_k\}$，并使用 ResNet 对检测到的视觉区域 r_i 进行编码，得到 $\boldsymbol{x}_i\in R^{2048}$，然后通过线性投影 $\boldsymbol{v}_i=\boldsymbol{W}_v\boldsymbol{x}_i+\boldsymbol{b}_i$，$\boldsymbol{W}_v\in\boldsymbol{R}^{d*2048}$，$\boldsymbol{b}_i\in\boldsymbol{R}^d$ 将 $\boldsymbol{x}_i$ 转换到和文本相同的向量中间中，从而得到最终图像 I 的细粒度表示为 $\boldsymbol{v}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k\}$，$\boldsymbol{v}_i\in\boldsymbol{R}^d$，并使用 $\boldsymbol{R}$ 的最大池化结果作为图像的全局表示 $\boldsymbol{I}^{\mathrm{glo}}$。

3.2 细粒度对齐

为了描述方面词和视觉区域之间的详细对应关系，并在不同的模态之间进行视觉语义对齐，本文使用标准化的基于距离的表示来体现异构模式之间的语义相似性。具体的图像区域 v_i 与方面词 $\boldsymbol{H}_{\mathrm{avg}}$之间的局部语义相似性 s_i^v 表示为：

$$s_i^v=\frac{\boldsymbol{W}_s^v\mid v_i-\boldsymbol{H}_{\mathrm{avg}}\mid^2}{\parallel\boldsymbol{W}_s^v\mid v_i-\boldsymbol{H}_{\mathrm{avg}}\mid^2\parallel_2} \tag{9}$$

其中，$\boldsymbol{W}_s^v\in\boldsymbol{R}^{p*d}$是一个可学习的参数矩阵，$p$ 为超参数，本文进一步测量了整个图像 $\boldsymbol{I}^{\mathrm{glo}}$和全文 $\boldsymbol{T}^{\mathrm{glo}}$之间的全局语义相似性 $\boldsymbol{S}^{\mathrm{glo}}$，同样 $\boldsymbol{W}^{\mathrm{glo}}\in R^{p*d}$为可学习的参数矩阵。

$$\boldsymbol{S}^{\mathrm{glo}}=\frac{\boldsymbol{W}^{\mathrm{glo}}\mid\boldsymbol{I}^{\mathrm{glo}}-\boldsymbol{T}^{\mathrm{glo}}\mid^2}{\parallel\boldsymbol{W}^{\mathrm{glo}}\mid\boldsymbol{I}^{\mathrm{glo}}-\boldsymbol{T}^{\mathrm{glo}}\mid^2\parallel_2} \tag{10}$$

本文使用全局语义相似度 $\boldsymbol{S}^{\mathrm{glo}}$和 $\boldsymbol{s}_i^v$ 之间的归一化相似性来测量匹配置信度 $\boldsymbol{c}_i$。

$$\xi_i=\boldsymbol{w}_i(\boldsymbol{S}^{\mathrm{glo}}\odot s_i^v), i=1,2,\cdots,k \tag{11}$$

$$\boldsymbol{c}=\sigma(LayerNorm([\xi_1,\xi_2,\cdots,\xi_k])) \tag{12}$$

其中，$\boldsymbol{c}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_k]$，$\boldsymbol{w}_i\in\boldsymbol{R}^{1*p}$为可学习参数向量，$\odot$表示两个向量对应元素相乘，$\sigma$ 为 sigmoid 激活函数，LayerNorm 表示归一化操作。该置信度的关键思想判断图像文本的整体语义相似度中包含了多少方面词和视觉区域之间的语义相似性，即该局部图像区域是否真的从图像文本全局的角度来对文本中方面词进行了描述。为了过滤不可靠视觉区域和方面词的相似性匹配，我们用每个视觉区域的相似度 $\boldsymbol{s}_i^v$ 乘以相应的置信度 $\boldsymbol{c}_i$。因此全局的语义相似性以及被置信度约束后的局部相似性为：

$$\boldsymbol{S}=[\boldsymbol{S}^{\mathrm{glo}},s_1^v\cdot c_1,s_2^v\cdot c_2,\cdots,s_k^v\cdot c_k] \tag{13}$$

然后，对 $\boldsymbol{S}$ 进行多层自注意力计算，以增强模态间的细粒度信息对齐。

$$S^{l+1}=\mathrm{ReLU}(\boldsymbol{W}_r^l\cdot\mathrm{softmax}(\boldsymbol{W}_q^lS^l\cdot(\boldsymbol{W}_k^lS^l)^T)\cdot S^l) \tag{14}$$

其中，$\boldsymbol{W}_q^l\in\boldsymbol{R}^{p*p}$和 $\boldsymbol{W}_k^l\in\boldsymbol{R}^{p*p}$分别是用来转换第 l 层的 query 和 key 的参数矩阵，$\boldsymbol{W}_r^l\in R^{p*p}$用来将输出维度映射到适合 $l+1$ 层输入的参数矩阵。然后将最后一层输出 $\boldsymbol{S}^L$ 的后 k 列按列做最大池化得到$\boldsymbol{\alpha}\in\boldsymbol{R}^k$，从 $\boldsymbol{\alpha}$ 中取最大值的下标 q，并取出相同下标的图片区域的特征表示 $\boldsymbol{v}_q$ 作为对齐模块的输出。

3.3 细粒度融合模块

该模块的作用是在方面词和视觉对象细粒度对齐的基础上，让方面词、文本、视觉对象、完整图片信息充分交互且协同互补，主要通过注意力机制来实现。

因为对方面词的上下文共同联合建模，对于提取相关的情感信息很重要，因此本文采用注意力机制来决定文本要更加关注的内容。同样将注意力机制应用于上阶段抽取出的 $\boldsymbol{v}_q$ 与图片完整信息 $\boldsymbol{R}$ 中，这有助于帮助模型只关注与视觉实体相关的可视块。其中注意力机制中计算相关性分数的计算式如下：

$$H_{asp}=\tanh(W_{asp}H_{asp}+b_{asp}) \tag{15}$$

$$\beta(H_{asp},H_i)=\tanh(H_{asp}W_{TA}H_i^T+b_{TA}) \tag{16}$$

$$\alpha_i^{TA}=\frac{\exp(\beta(H_{asp},H_i))}{\sum_{j=1}^{49}\exp(\beta(H_{asp},H_j))} \tag{17}$$

其中，$\boldsymbol{W}_{asp}\in\boldsymbol{R}^{d*d}$，$\boldsymbol{b}_{asp}\in\boldsymbol{R}^d$，$\boldsymbol{W}_{TA}\in\boldsymbol{R}^{d*d}$，$\boldsymbol{b}_{TA}\in\boldsymbol{R}$ 为可训练参数，得到注意力分数向量 $\boldsymbol{\alpha}^{TA}\in\boldsymbol{R}^n$，同样地，$\boldsymbol{\alpha}^{IO}\in\boldsymbol{R}^k$由 $\boldsymbol{v}_q$ 与 $\boldsymbol{R}$ 计算得到。基于 $\boldsymbol{\alpha}^{TA}$，$\boldsymbol{\alpha}^{IO}$可以得到最终的文本、视觉上下文表示 H_T，H_I。

$$H_T=\sum_{i=1}^{n}\alpha_i^{TA}H_i \tag{18}$$

$$H_I=\sum_{i=1}^{49}\alpha_i^{IO}R_i \tag{19}$$

为了更好地实现方面词与图像、视觉实体与文本两两之间的跨模态细粒度交互，虽然许多先进的方法都使用简单的特征串联，但我们认为，这样会忽略他们之间的高阶交互作用，因此本文首先使用文献[27]中提出的多头交互注意力机制来计算跨模态交互信息。

$$\begin{aligned}CATT^i&=(Q,K,V)\\&=\mathrm{softmax}\left(\frac{[\boldsymbol{W}_Q^iQ]^T[\boldsymbol{W}_K^iK]}{\sqrt{\frac{d}{m}}}\right)[\boldsymbol{W}_V^iV]^T\end{aligned} \tag{20}$$

$$\ddot{C}=\boldsymbol{W}_m[CATT^1(Q,K,V),\cdots,CATT^m(Q,K,V)]^T \tag{21}$$

$$\boldsymbol{H}_{TO}=CATT(v_q,H,H) \tag{22}$$

$$\boldsymbol{H}_{IA}=CATT(H_{asp},R,R) \tag{23}$$

其中，m 为交互注意力头的数量，$\{\boldsymbol{W}_Q^i,\boldsymbol{W}_K^i,\boldsymbol{W}_V^i\}\in\boldsymbol{R}^{d/m*d}$是 query，key 和 value 的权重矩阵，$\boldsymbol{W}_m\in\boldsymbol{R}^{d*d}$是多头交互注意力机制的参数矩阵，计算出的 $\boldsymbol{H}_{TO}$，$\boldsymbol{H}_{IA}$为方面词与图像、视觉实体与文本两两之间的跨模态细粒度交互信息。

然后采用文献[1]提出的低秩双线性池对 $\boldsymbol{H}_{TO}$，$\boldsymbol{H}_{IA}$进行融合，可以保证在用更少的参数情况下保持标准双线性运算符的性能，其计算过程如下。

$$\boldsymbol{H}_M=W_M(\sigma(\boldsymbol{W}_1\boldsymbol{H}_{TO})\circ\sigma(\boldsymbol{W}_2\boldsymbol{H}_{IA}))+b_M \tag{24}$$

其中，$\boldsymbol{W}_M$，$\boldsymbol{W}_1$，$\boldsymbol{W}_2\in\boldsymbol{R}^{d*d}$，$\boldsymbol{b}_M\in\boldsymbol{R}^d$ 均为可训练参数，σ 是非线性变换函数 tanh 函数，$\circ$为按元素相乘。将 $\boldsymbol{H}_I$，$\boldsymbol{H}_T$，$\boldsymbol{H}_M$ 结合得到最终的多模态表示：

$$\boldsymbol{H}_{\mathrm{final}}=[\boldsymbol{H}_I\mid\boldsymbol{H}_T\mid\boldsymbol{H}_M] \tag{25}$$

3.4 情感标签分类

将融合后的多模态向量表示 $\boldsymbol{H}_{\mathrm{final}}$送入 softmax 中，用于方面级情感分类，取输出中概率最高的标签作为最终结果。其中 $\boldsymbol{W}\in\boldsymbol{R}^{3b*3}$，$\boldsymbol{b}\in\boldsymbol{R}^3$ 为可学习参数。

$$\hat{y}=\mathrm{softmax}(\boldsymbol{W}^T\boldsymbol{H}_{\mathrm{final}}+b) \tag{26}$$

3.5 模型训练

为了优化本文模型中所有的参数，本文的目标是最小化交叉熵损失函数。

$$Loss=-\sum_{i=1}^{3} y^i \log(\hat{y}^i) \tag{27}$$

4 实验

4.1 数据集与实验参数

实验主要采用文献[1]中的两个基准数据集 TWITTER-2015 和 TWITTER-2017 来评估本文模型，主要包含 2014－2015 年和 2016－2017 年发布的多模态用户帖子，所有方面词实体都属于 4 种类别：人、地点、组织和其他。其中包含文本和与之对应的图片，并且标注了目标方面词以及图文对该方面词的情感倾向，情感标注为三分类，两个数据集的统计数据如表 1 所列。将数据集按照 3∶1∶1 的比例划分为训练集、验证集、测试集，表 2 列出了 3 种数据集的情感标签分布。模型的参数设置如表 3 所列，模型使用 xavier_uniform_优化器来更新模型参数，整个模型采用 python 语言，pytorch 框架。本文与很多细粒度情感分析的文章相似，使用 F1 分数和准确率来衡量模型的性能。早停轮数指在训练模型的过程中，验证集 F1 分数在连续几个 epoch 中都未取得提升，从而停止模型训练，防止过拟合。

表 1 实验数据统计

Table 1 Experimental data statistics

Attribute	Twitter2015	Twitter2017
Tota lnum	3179	3562
Max Length	13.2	13.9
Min Length	24	27
Avg Aspect	1.6	2.2

表 2 数据集情感标签分布

Table 2 Distribution of sentiment labels in datasets

	Twitter2015				Twitter2017			
	Pos	Neg	Neu	Total	Pos	Neg	Neu	Total
Train	928	368	1883	3179	1508	416	1638	3562
Dev	303	149	670	1122	515	144	517	1176
Test	317	143	607	1037	493	168	573	1234

表 3 模型参数

Table 3 Model parameters

参数名称	参数数值
学习率	0.001
迭代次数	8
批处理大小	10
文本嵌入维度	100
图片特征维度	2048
早停轮数	4

4.2 对比实验

本节将 TFGA 模型与现有的 TMSC 模型的几种代表性方法进行了比较。表 4 列出了每种方法的准确性(acc)和 F1 分数。本文首先考虑了以下只关注文本的方法并对其进行了比较。

1)RAM[28]。使用 Bi-LSTM 学习上下文的隐藏表示产生 Memory，然后给 Memory 进行位置加权，权重与各单词和方面词的相对距离有关，使得来自同一个句子的不同的方面词有了量身定做的 Memory。最终将多注意力机制用于位置加权后的 Memory，再将输出结果和 RNN 结果非线性结合，得到最终的上下文表示。

2)TD-LSTM/ TC-LSTM[29]。将文本以方面词为断点分为左右两部分，使用 LSTM 分别对其进行建模，然后用最后一个时间步的隐藏状态向量拼接起来，得到最终的上下文表示，进行情感分类。

3)MGAN[11]。使用双向 LSTM 并引入位置编码获得方面词和上下文的语义表示，并设计了一种粗细粒度结合的注意力机制，分别计算 aspect 中每个词和上下文中的每个词之间的双向影响(细粒度)，上下文的平均语义表示给 aspect 的每个词分配不同的权重，aspect 的平均语义表示给所有上下文词分配不同的权重(粗粒度)，然后将根据权重调整后的方面词与上下文语义表示进行拼接，再进行情感分类。

4)ESTR[1]。分别使用 3 个 LSTM 来获取左上下文、右上下文以及方面词中每个词的隐藏状态，然后利用方面词给每一个上下文词生成适当的注意力权重，左右上下文词的加权和分别被视为左右上下文的文本表示，然后通过双线性融合层融合方面词信息，从而获得最终的文本表示。

此外，我们还考虑了多模态方法并对其进行了比较。

1)Res-MGAN/Res-RAM/ Res＋aspect。对视觉特征做最大池化，获得 $\boldsymbol{g}=\text{MaxPooling}(\text{ResNet}(\boldsymbol{I}))$，然后分别与 MGAN，RAN 的文本表示或者方面词的向量表示直接拼接，将结果送入 softmax 中进行分类。

2)MIMN[19]。对于上下文、方面词和图片分别采用双向 LSTM 和 CNN 来获取隐藏表示，然后使用多层记忆网络对方面词、上下文和视觉上下文之间的交互进行建模，从而融合两个模态的信息，进行情感分类。

3)TomBERT[21]。一种多模态 bert 架构，采用 bert 获取基于方面词的文本表示，并设计了一种目标注意力机制，做方面词图像匹配，获得面向方面词的视觉表示，并在顶部堆积了自注意力机制，捕获多模态之间的交互信息。

4)TomLSTM。TomBERT 的变形，即将 TomBERT 模型中的 target encoding 和 sentenceencoding 替换成 LSTM，来获得文本的特征向量。

5) ESAFN[1]。在 ESTR 获得文本表示的基础上，使用 Resnet-152 获得图片的每个视觉块的特征向量，然后根据每块和方面词的相关性计算每块的关注权重，从而获得面向实体的视觉表示，并采用视觉门控机制来消除无关文本的视觉信息带来的噪音。最后使用双线性融合层聚合文本和图片信息得到多模态表示，再将其输入 softmax 函数中进行实体级情感分类。

6)TomLSTM＋align：在 TomLSTM 的方面词注意力后加一个 softmax 层，以获得视觉块的相关性分数，并使用硬注意力机制进行筛选，然后进行情感分析。

4.2.1 主要对比实验

表 4 列出了本文模型与各基线模型的对比结果。为了避免模型训练过程中的随机性，本文所有实验均进行了 5 次，并取 5 次结果的平均值，进而更加客观地对模型结果进行描述。观察实验结果可以发现，在两个数据集上，本文 TFGA 模型在 ACC 和 F1 两个指标上都优于绝大部分基线模型。这是由于 TFGA 模型对文本和图片进行细粒度对齐，并将文本、

方面词、图片、可视对象进行了充分交互融合，弱化了图像中噪声信息对模型的影响，从而提取出有用的关键信息。TD-LSTM 模型将文本方面词的上下文分开建模的性能非常有限，这表明方面词的局部上下文对情感分析的综合影响不应忽视。由于视觉模态的加入，模型性能得到了一定的改进，这说明图像确实可以对文本起到支持作用，提供补充信息。而 Res-aspect 模型的效果欠佳，主要是因为上下文信息没有得到很好的利用。另外，可以观察到 TomBERT 模性的性能优于变形的 TomLSTM，这是合理的，因为 TomBERT 采用了预训练语言模型，其特征提取能力优于 LSTM。MIMN 模型在图像和文本信息两者中使用注意力机制建模文本与图像之间的交互，其性能优于大部分模型，但 MIMN 使用完整图像信息与文本信息相融合，作为最终方面词的向量表示 $\boldsymbol{H}_{\text{avg}}$。基于这些隐藏状态，进一步采用广泛使用的注意力机制[18]并引入了图片中的噪声信息，因此它的性能逊于本文模型，这充分说明了本文模型中的细粒度对齐的必要性。

表 4 对比实验结果

Table 4 Results of contrast experiment

Modality	Method	Twitter2015		Twitter2017	
		ACC	*F*1	*ACC*	*F*1
Text	RAM	70.54±0.20	63.15±0.28	64.42±0.37	61.01±0.31
	ESTR	71.06±0.26	64.28±0.36	65.64±0.23	62.07±0.36
	MGAN	71.09±0.29	64.21±0.24	64.35±0.31	61.26±0.28
Text+vision	Res-aspect	59.48±0.37	46.28±0.31	58.64±0.32	53.45±0.40
	Res_RAM	71.25±0.33	64.58±0.30	65.21±0.37	62.03±0.27
	Res_MGAN	71.35±0.36	63.66±0.27	66.27±0.26	63.34±0.22
	TomLSTM	73.30±0.37	67.47±0.31	67.63±0.28	64.42±0.23
	TomBERT	75.96±0.22	71.09±0.23	70.02±0.38	67.89±0.21
	MIMN	69.54±0.31	63.49±0.28	64.66±0.38	61.51±0.36
	ESAFN	73.07±0.36	67.17±0.29	67.63±0.26	64.26±0.23
	TFGA	**75.33±0.15**	**70.31±0.20**	**70.47±0.20**	**66.38±0.25**

4.2.2 细粒度对齐效果实验

本文使用 TomLSTM，TomLSTM+align 和 TFGA 在文献[30]提出的从 Twitter2017 中随机选择的图像目标匹配数据集上的情感分类结果来测试细粒度对齐的实验效果。实验结果如表 5 所列，首先 TomLSTM+align 对齐比 TomLSTM 的效果差，本文推测其原因是，使用 Resnet 获得视觉特征中包含较少的视觉目标信息，并且会给对齐过程带来一定的噪声。其次结果显示 TFGA 模型优于其他两个模型，说明本文提出的细粒度对齐机制在视觉区域和方面词对齐方面较有优势，且在 MABSA 任务中具有一定的重要性。

表 5 细粒度对齐对比实验结果

Table 5 Results of contrast experiment of fine-grained alignment

Method	*ACC*(2017)
TomLSTM	73.39
TomLSTM+align	72.45
TFGA	75.33

4.2.3 参数 k 的影响实验

对于本文模型，通过从图片中提取不同个数的视觉区域来比较模型的性能，如图 3 所示，随着 k 值变大，模型的准确率不断提升，当 $k=8$ 时到达顶峰，然后随着 k 再次变大，准确率逐渐降低。因为所使用的数据集中大部分样本所包含的方面词不多于 4 个，当 k 值过多时会引入噪声，因此性能变差。

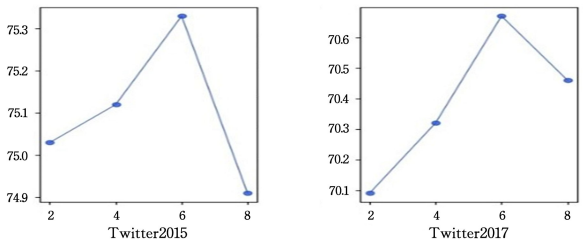


图 3 参数 k 对模型性能的影响

Fig. 3 Influence of parameter k on model performance

4.3 消融实验

表 6 列出了本文的消融实验结果。可以看出，删除第 3.1.2 节中的视觉表示模块，直接使用完整的图像特征与文本模态进行细粒度对齐与融合的性能相比 TFGA 模型显著降低，这证明了使用 FasterRCNN 获得细粒度的图像表示来进行对齐与融合的重要性。为了证明置信度约束对模型性能的影响，本文进行了没有置信度约束的消融实验，模型的准确率降低了 1.1%，在这种情况下，方面词与相似性的不同视觉区域之间的局部语义相似性差别不大，混淆了模型的视线导致其性能下降，因此证明了置信度约束可以帮助模型过滤掉不可信的视觉区域。此外，还可以观察到，将对齐机制中的硬注意力机替换为相关性较高的前 6 个可视区域特征的相关性加权和的方式，对模型性能有一定程度的影响，这说明使用对齐机制过滤掉具有噪声的视觉信息对模型是有用的。去除细粒度融合机制后，模型采用拼接融合的方式，实验结果表明对文本和图片的整体信息、方面词以及视觉区域之间进行细粒度交互能够帮助模型获取更准确的用于方面词情感分析的特征。

表 6 消融实验结果

Table 6 Ablation experiment results

Approaches	Twitter2015		Twitter2017	
	ACC	*F*1	*ACC*	*F*1
TFGA	75.33	70.31	70.47	66.38
w/o FastRCNN	69.01	62.37	64.38	61.47
w/o 置信度	74.23	67.42	66.96	63.71
w/o 硬注意力机制	75.21	68.24	68.25	65.38
w/o 细粒度交互融合	75.01	68.16	67.97	65.12

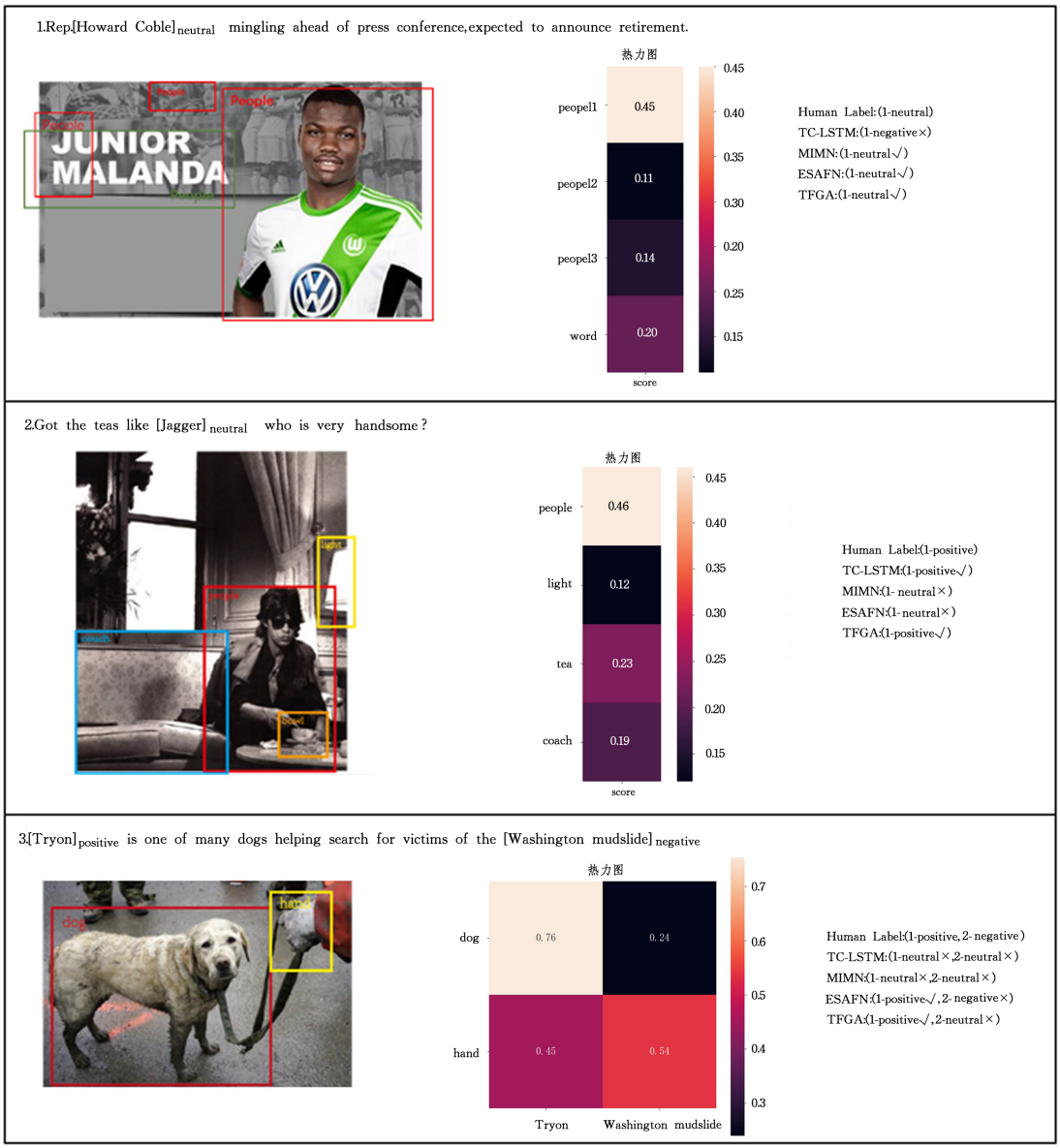
4.4 案例研究

为了证明本文模型的优势，本文对两个数据集中的测试集进行了额外测试，并仔细挑选了几个具有代表性的测试样本，用于分析模型的预测性能。表 7 列出了基线方法和本文 FGAF 模型对案例样本的预测之间的比较。首先在表 7 的

第一项，图片中出现了多个类别为 people 的视觉实体，在置信度分数的约束下，模型成功地聚焦到相关性以及可信度都较高的 people1 数据区域，并借助该视觉区域中握手的相关图像信息，纠正了纯文本模型的错误预测。在表 7 的第二项中，由于方面词下文表达出了强烈的积极情绪，因此纯文本模型给出了正确的预测，但是对于多模态模型，我们可以看到与方面词 Jagger 对应的图像区域在右下角的位置，ESAFN 和 MIMN 模型都使用了完整的图像特征与文本模型进行融合，引入了图像无关区域的噪声信息，因此给出了错误预测，但本文的 TFGA 模型可以识别到图像中与 Jagger 相关性较高的区域，因此正确地预测出实体情绪为积极。类似地，在表中的第三项中，在文本具有多个方面词的情况下，TFGA 模型可以更加准确地聚焦到图与每一个方面词相关的图片区域上，从而做出正确的预测。

表 7 实验样例

Table 7 Experimental examples



4.5 错误分析

为了分析本文模型的局限性，我们从两个数据集的测试集中都随机选择了 150 个预测错误的示例，并分析了其中 3 类标签的分布情况，发现大部分与中性类别相关（Twitter2015 占比 67.3%，Twitter2017 占比 69.4%），其中图像中包含了与方面词相关性较少的信息，并且其中图片中不同视觉区域与方面词的相关性以及其对应的置信度分数的分布都较为均匀。本文认为这是因为用户更倾向于将自己感兴趣或者讨厌的这两类感情色彩较为浓烈的信息用图片的方式表达出来，因此对于这两类样本，模型更容易在图片信息的帮助下进行正确的预测。相反，中性类别的方面词缺乏相应的情感表达，模型受到噪声的影响会更大。图 4 给出了一个典型的错误案例，方面词 Rwanda 为中性情感，而图片中描述的人类的面部信息带有较多积极情感的信息，因此误导模型判断出正向情感。而对于除中性类别外的错误，大多是因为图像和文本的关系不密切，例如在表 7 的第三项中，我们可以明显观察到对于方面词 Washington mudslide 的情绪应该是消极的，但由于图片中并没有与华盛顿泥石流相关的区域，被抽取出来的视

觉区域与该方面词的相关性分数以及对应的置信度分数经过归一化后相差并不大，因此在这种情况下筛选出来的视觉区域对于方面词的情感极性的判断并没有正向的帮助，反而带了一定噪声信息，导致模型预测出了错误的结果。

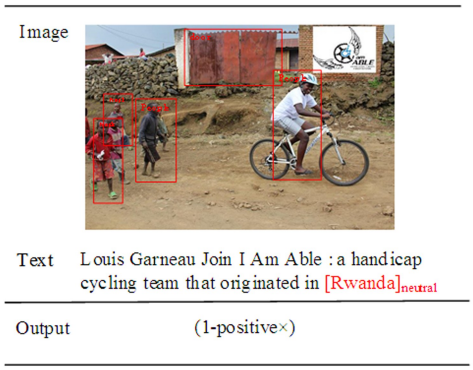


图 4 错误案列

Fig. 4 Error case

结束语 本文提出了一种可信细粒度对齐的多模态方面级情感分析模型，对于文本数据集，本文采用双向 LSTM 对上下文共同建模，从而学习文本的上下文表示，很好地解决了文本序列中的长期依赖问题。对于图片数据，本文采用预训练好的 FasterRCNN 模型抽取出文本中的视觉对象，在置信度的约束下计算视觉对象与方面词之间的局部语义相似性，并通过硬注意力机制进行筛选，选择出图片中与方面词相关性最高的视觉区域，并对两个模态信息进行细粒度融合，然后得到最终的情感分类。

本文未使用效果更好的预训练语言模型，后续将对此做进一步改进。由于本文使用的基准数据集本身带有方面词，因此会限制实际应用。接下来将继续探索使用端到端的方式来联合抽取方面词和情感极性，并融入细粒度对齐，从而提升任务的准确率。

参 考 文 献

[1] 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, 2019, 28: 429-439.

[2] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2017, 39(6): 1137-1149.

[3] LIU Q, LI N, TIAN Y A. Annotation of Logical Structure in Reflowable Document for Machine Learning[J]. Journal of Chinese Information Processing, 2019, 33(9): 50-59, 78.

[4] REXIDANMU T, WUSHOUR S, YIERXIATI T. Uyghur Text Sentiment Analysis by Combining Lexical Knowledge with Machine Learning Methods[J]. Journal of Chinese Information Processing, 2017, 31(1): 177-183.

[5] 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.

[6] TAY Y, TUAN L A, HUI S C. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[7] NGUYEN H T, LE NGUYEN M. Effective attention networks for aspect-level sentiment classification[C]// 2018 10th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2018: 25-30.

[8] LIU J, ZHANG Y. Attention modeling for targeted sentiment [C] // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 572-577.

[9] CHENG J, ZHAO S, ZHANG J, et al. Aspect-level sentiment classification with heat (hierarchical attention) network[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 97-106.

[10] MA D H, LI S J, ZHA X D, et al. Interactive attention networks for aspect-level sentiment classification[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: International Joint Conferences on Artificial Intelligence, 2017: 4068-4074.

[11] 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.

[12] XUE W, LI T. Aspect based sentiment analysis with gated convolutional networks[J]. arXiv: 1805. 07043, 2018.

[13] LI X, BING L, LAM W, et al. Transformation networks for target-oriented sentiment classification [J]. arXiv: 1805. 01086, 2018.

[14] ZHANG M, ZHANG Y, VO D T. Gated neural networks for targeted sentiment analysis[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[15] DAI J, YAN H, SUN T, et al. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta[J]. arXiv: 2104. 04986, 2021.

[16] SUN C, HUANG L, QIU X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence[J]. arXiv: 1903. 09588, 2019.

[17] WU Z, ONG D C. Context-guided bert for targeted aspect-based sentiment analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 14094-14102.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neutral Information Processing. Red Hook, NY: Curran Associates Inc, 2017: 5998-6008.

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

[20] LIU L L, YANG Y, WANG J. ABAFN: Aspect-Based Sentiment Analysis Model for Multimodal [J]. Computer Engineering and Applications, 2022, 58(10): 193-199.

[21] YU J, JIANG J. Adapting BERT for target-oriented multimodal

sentiment classification[C]// IJCAI. 2019.
[22] YU Y, ZHANG D, LI S. Unified Multi-modal Pre-training for Few-shot Sentiment Analysis with Prompt-based Learning [C]// Proceedings of the 30th ACM International Conference on Multimedia. 2022: 189-198.
[23] 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.
[24] 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.
[25] YU J, CHEN K, XIA R. Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis [J/OL]. IEEE Transactions on Affective Computing, 2022. https://newsletter. x-mol. com/paper/1534586027781197824.
[26] JU X C, ZHANG D, XIAO R, et al. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 4395-4405.
[27] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences [J]. arXiv: 1906. 00295, 2019.
[28] CHEN P, SUN Z, BING L, et al. Recurrent attention network on memory for aspect sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 452-461.
[29] KIRITCHENKO S, ZHU X, CHERRY C, et al. Detecting aspects and sentiment in customer reviews[C]// 8th International Workshop on Semantic Evaluation (SemEval). 2014.
[30] YU J, WANG J, XIA R, et al. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022). 2022: 4482-4488.

FAN Dongxu, born in 2000, postgraduate. Her main research interests include sentiment analysis and data mining.

GUO Yi, born in 1975, Ph.D, professor. His main research interests include text mining, knowledge discovery and business intelligence.

（责任编辑：喻藜）