智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

采用目标注意力的方面级多模态情感分析研究

朱超杰, 闫昱名, 初宝昌, 李刚, 黄河燕, 高小燕

引用本文:
朱超杰, 闫昱名, 初宝昌, 等. 采用目标注意力的方面级多模态情感分析研究[J]. 智能系统学报, 2024, 19(6): 1562-1572.
ZHU Chaojie, YAN Yuming, CHU Baochang, et al. Aspect-level multimodal sentiment analysis via object-attention[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1562-1572.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202404009

DOI: 10.11992/tis.202404009
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240918.1650.002

采用目标注意力的方面级多模态情感分析研究

朱超杰[1], 闫昱名[2], 初宝昌[2], 李刚[2], 黄河燕[1], 高小燕[3]
（1. 北京理工大学 计算机学院, 北京 100081; 2. 北京华电电子商务科技有限公司, 北京 100073; 3. 北京工业大学 计算机学院, 北京 100124）

摘　要: 方面级的多模态情感分析（aspect-level multimodal sentiment analysis, ALMSA）旨在识别出语句和图像信息在某个特定方面上所表现出的情感极性。该任务现有分析模型使用的均是图像的全局特征, 并未考虑原始图像信息中的细节信息。针对这一问题, 提出一种基于目标注意力的方面级多模态情感分析模型 OAB-ALMSA（object-attention based aspect-level multimodal sentiment analysis）。采用目标检测算法捕获原始图像中目标的细节信息; 引入目标注意力机制并构建迭代的融合层来完成多模态信息的充分融合; 针对数据较高的复杂性所导致的训练困难问题, 为模型制定课程式学习策略。经课程式学习训练的 OAB-ALMSA 模型在 TWITTER-2015 数据集上得到了最高的 F_1, 这表明对图像中细节信息的利用能够提高模型对数据的综合理解, 提升预测效果。

关键词: 方面级情感分析; 多模态; 情感分析; 目标检测; 自注意力机制; 自然语言处理; 深度学习; 特征提取
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)06−1562−11

中文引用格式: 朱超杰, 闫昱名, 初宝昌, 等. 采用目标注意力的方面级多模态情感分析研究 [J]. 智能系统学报, 2024, 19(6): 1562–1572.
英文引用格式: ZHU Chaojie, YAN Yuming, CHU Baochang, et al. Aspect-level multimodal sentiment analysis via object-attention[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1562–1572.

Aspect-level multimodal sentiment analysis via object-attention

ZHU Chaojie[1], YAN Yuming[2], CHU Baochang[2], LI Gang[2], HUANG Heyan[1], GAO Xiaoyan[3]
(1. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China; 2. Beijing Huadian E-Commerce Technology Co., Ltd., Beijing 100073, China; 3. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Aspect-level multimodal sentiment analysis (ALMSA) aims to identify the sentiment polarity of a specific aspect word using both sentence and image data. Current models often rely on the global features of images, overlooking the details in the original image. To address this issue, we propose an object attention-based aspect-level multimodal sentiment analysis model (OAB-ALMSA). This model first employs an object detection algorithm to capture the detailed information of the objects from the original image. It then applies an object-attention mechanism and builds an iterative fusion layer to fully fuse the multimodal information. Finally, a curriculum learning strategy is developed to tackle the challenges of training with complex samples. Experiments conducted on TWITTER-2015 data sets demonstrate that OAB-ALMSA, when combined with curriculum learning, achieves the highest F_1. These results highlight that leveraging detailed image data enhances the model's overall understanding and improves prediction accuracy.

Keywords: aspect-level sentiment analysis; multimodal; sentiment analysis; object detection; self-attention; natural language processing systems; deep learning; feature extraction

情感分析是自然语言处理中的一个重要任务, 近些年来在商业智能、社交媒体、公共管理等方面都有亮眼表现, 因此也得到了越来越多的关

收稿日期: 2024−04−11.　**网络出版日期:** 2024−09−19.
基金项目: 国家自然科学基金项目（U21B2009）; 横向科技项目（2023110051000823）.
通信作者: 高小燕. E-mail: gaoxiaoyan@bjut.edu.cn.

注和发展。情感分析的目的是利用自然语言处理技术对带有主观性的文本信息进行分析、处理, 以此来提取其中所包含的情感色彩。近些年, 互联网高速普及, 据中国互联网信息中心发布的报告显示[1], 截止 2022 年 12 月我国网民数量已达到了总人口的 76.2%。此外, 网络技术的高速发展, 也使得人们在社交网络上表达情感的方式产生了许多变化, 为了应对这些变化, 情感分析发展出了许多不同的研究方向, 其中就包括方面级的多模态情感分析。

方面级的多模态情感分析作为情感分析任务中的一个重要分支, 其任务是根据给定的语句和图像信息, 通过分析得到整体数据在语句中某个特定的方面词上所表现出的情感极性。与传统的情感分析任务相比, 该任务旨在从复杂信息中挖掘出与某一个具体方面相关的细粒度信息, 并对数据在该方面所表现出的情感极性进行判断。因此, 方面级的多模态情感分析任务在如今这个网络信息越来越复杂的时代可以发挥更大的作用。

近些年来, 研究人员在方面级的多模态情感分析领域取得了很多研究成果, 其中有几个特点较为突出的研究模型: Xu 等[2] 基于注意力机制构建了记忆网络来完成多模态信息之间的融合, Zhang 等[3] 在使用注意力机制来完成多模态信息融合的过程中, 构建了融合判别矩阵对其进行监督, 从而提高了信息融合的效果; Khan 等[4] 使用预训练模型 Transformer 将图像转化为描述性的语句, 将多模态任务转化为单模态的任务。但是, 目前所进行的方面级多模态情感分析工作主要关注于使用图像的全局特征信息进行多模态的信息融合, 却忽视了原始图像中的细节信息。

为了解决上述问题, 本文引入目标检测算法并构建目标注意力机制来对原始图像中的目标信息进行检测和利用, 并以此为基础提出了基于目标注意力的方面级多模态情感分析模型 OAB-ALMSA(object-attention based aspect-level multimodal sentiment analysis)。针对数据集复杂度高导致训练困难, 本文为该模型构建了相对应的课程式学习训练策略。为了评估该方法, 本文在数据集 TWITTER-2015 上进行了实验。实验结果表明, 本文方法在性能上明显优于对比方法。本文主要贡献如下:

1) 提出一个基于目标注意力的方面级多模态情感分析模型 OAB-ALMSA。该模型通过引入目标注意力机制, 构建多模态融合模型, 实现了图像的目标检测结果与文本信息之间的融合, 从而提取有利信息来完成情感分析任务。

2) 采用课程式学习方法, 对 OAB-ALMSA 模型进行训练。该方法通过对模型进行预训练来获取多模态信息的向量化表示, 以此来完成样本评价和数据集划分, 进而设计训练过程并完成模型的再训练。

3) 本文在方面级多模态情感分析任务的公开基准数据集 TWITTER-2015 上进行了实验, 通过与相关模型的对比发现, 经课程式训练的 OAB-ALMSA 模型有明显的性能提升, 这证明了目标注意力机制以及课程式训练策略的有效性。

1　相关工作

1.1　方面级文本情感分析

传统的方面级情感分析任务主要是基于纯文本信息进行的, 在目前的自然语言处理领域已经取得了较为丰富的研究成果。方面级的文本情感分析任务主要包含 3 种不同的方法: 基于情感词典的方法、基于传统机器学习的方法和基于深度学习的方法。

基于情感词典的方法主要侧重于根据情感词典所提供情感词的极性来对需要分析的语句进行情感极性划分, 因此, 该方法最为关键的一步就是构建成熟的情感词典。王伟贤等[5] 基于现有的库构建了基础情感词典, 并基于统计信息识别新词, 从而构建了完整的微博情感词典。然而该类方法需要大量的先验知识来构建词典, 这不仅会耗费大量的人力, 在如今这个信息高速发展的时代, 其推广能力也显得不足。在此之后, 相关的研究主要着重于使用机器学习方法来完成情感分析任务[6]。

基于传统机器学习方法主要通过对支持向量机、隐马尔可夫、朴素贝叶斯等模型的训练来完成情感分类。Huang 等[7] 提出了用于联合方面情感主题嵌入的 JASen 模型, 该模型在所有联合主题上对用户给定关键词的联合分布进行建模来使得主题嵌入周围的单词可以很好地描述联合主题的语义。但这一类方法对标注数据的数量和质量都有着较高的需求[8], 因此需要投入庞大的人力物力。此外, 情感分析需要考虑多个层面的特征, 而这些特征可能是高度稀疏的[9], 这使得机器学习模型很难对这些特征进行很好的学习。

近些年来, 深度学习技术的发展和使用在一定程度上改善了这些问题。研究人员通过结合不同类型的神经网络来完成模型的构建, 发展出了许多不同的方法。如: 基于循环神经网络(recurrent neural network, RNN)[10]、基于卷积神经网络

(convolutional neural network, CNN)[11-12]、基于门控网络的方法[11]。此外,受到注意力机制在自然语言处理任务中所表现出优势的启发,许多研究人员开始进行基于注意力机制模型的设计。Wang 等[13]提出了一个基于分割注意力的长短期记忆模型,该模型通过条件随机场来有效地捕捉情感表达与方面之间的结构相关性。曾锋等[14]使用双层注意力分别对单词层和句子层进行建模,以此获取深层次的情感特征信息。谢珺等[15]通过两种注意力机制对上下文的语义和句法结构进行信息融合,以获得更丰富的信息。李丽双等[16]通过使用 Bi-GRU 和多注意力机制提取情感信息,使其与目标表现完美融合。而近几年来,预训练的语言模型因其无需人工标签,并可以从海量的语料中学习到通用的语言表示,进而显著优化下游任务的结果,开始逐渐成为研究的主流。Wu 等[17]通过使用具有上下文感知能力的 Transformer 来生成上下文引导的 BERT(bidirectional encoder representation from transformers) 模型 CG-BERT(context-guided BERT), 进而完成了情感分析任务。曾凡旭等[18]认为传统模型所使用的 Word2vec 和 GloVe 等词向量并不能捕获到语义的上下文相关性,因此选择使用预训练的语言模型 BERT 来自动学习带有上下文信息的特征表示。

虽然多年的发展使得纯文本的方面级情感分析取得了显著的成功,但是单纯的对文本信息进行处理已经无法有效地对当今社交媒体多样的数据进行分析,因此方面级多模态情感分析任务逐渐走进了科研人员的视野。

1.2　方面级多模态情感分析

近些年来,由于互联网及其技术的高速发展,人们在网络上表达自身情感的方式越来越多样化和复杂化,因此,方面级多模态情感分析任务开始受到越来越多的关注。方面级的多模态情感分析任务于 2019 年由 Xu 等[2]提出,经过多年的发展,研究人员在基于深度学习的方法上取得了多种多样的成果。

与方面级的文本情感分析类似,方面级的多模态情感分析在发展过程中同样受到了注意力机制在自然语言处理任务中优异表现的影响,因此,大多数模型在构建的过程中都包含有注意力机制结构。Gu 等[19]认识到普通的注意力机制在处理信息的过程中只能关注信息的某一方面,因此选择在构建模型时使用多头注意力机制来分别完成方面词和语句、方面词和图像之间的信息融合。Xu 等[2]选择基于注意力机制来构建记忆模块进而完成模态间信息的融合,以提高模型在处理复杂的图像信息和长语句信息时的表现。范东旭等[20]选择使用自注意力机制对图像的全局特征进行处理,以获取图像内的细粒度信息来完成多模态信息融合。

在使用注意力机制等结构来进行多模态信息融合的过程中,许多研究人员选择引入额外的方法来辅助模型的训练。如 Zhou 等[21]在进行方面级的多模态情感分析任务时引入了对抗训练(adversarial training)的思想。通过对抗训练来辅助文本和图像的对齐,模型可以更好地完成多模态之间的信息融合。Zhang 等[3]选择构建了一个融合判别矩阵来监督多模态信息之间的融合过程,这使得模型可以识别不同模态之间的一致性和冗余性,进而提高模型情感预测的准确性。

除此之外,BERT 等预训练模型的出现也启发了相关工作的研究人员。Yu 等[22]通过对预训练语言模型 BERT 中的注意力机制进行调整来获取文本信息与图像信息之间的交互,构建了名为 TomBERT 的方面级多模态情感分析模型。Khan 等[4]则是选择利用 Transformer 模型将数据中的图像信息全部转化为了描述性的文本信息,与原本的图像相比,文本更加简练,信息密度更大,同时便于处理,因此可以极大地降低任务的难度。Ling 等[23]基于生成式的预训练模型 BART(bidirectional and auto-regressive transformers)为方面级多模态情感分析任务设计了一个特定任务的语句–图像预训练框架。

然而,在多模态信息融合过程中,研究人员使用的主要为图像的全局信息,并没有考虑原始图像中所存在的细节信息。模型的训练过程也没有考虑到该任务下数据集的复杂性所导致的模型收敛困难的问题。基于此,本文提出 OAB-ALMSA 模型以及对应的课程式训练策略。该模型通过目标检测算法获取原始图像中所存在的目标,将其作为图像的细节信息,并引入目标注意力机制来完成目标细节信息与方面词、语句之间充分的信息交互,从而提高情感分析的性能。最后,我们对模型进行预训练,并根据得到的模型参数对样本数据进行处理,进而设计课程式训练策略,并完成对模型的再训练。

2　OAB-ALMSA 模型

OAB-ALMSA 模型主要由 5 个部分组成,嵌入层、方面级的单模态融合层、跨模态融合层、迭代层和情感预测层,除此之外,本文还通过构建

课程式学习策略对模型进行优化, 构建了性能更强的 OAB-ALMSA-CL(OAB-ALMSA via curriculum learning)模型。模型 OAB-ALMSA 的整体结构如图 1 所示。该模型旨在对一个含有方面词的多模态数据集 $\boldsymbol{D}$ 进行分析, 生成一个序列 $\boldsymbol{Y}$ 作为情感预测结果。数据集 $\boldsymbol{D}$ 中的每一个元素 $\boldsymbol{d}$ 包含一个图像 $\boldsymbol{I}$ 以及与该图像对应的包含有 o 个单词的语句 $\boldsymbol{S}=\{\omega_1,\omega_2,\cdots,\omega_o\}$, 以及一个包含有 m 个单词的方面词 $\boldsymbol{A}=\{\omega_{a+1},\omega_{a+2},\cdots,\omega_{a+m}\}$, 并且该方面词是语句 $\boldsymbol{S}$ 的子句, 其中 $a+1$ 为该方面词起始单词在语句中的位置。对于数据集中的每一个元素 $\boldsymbol{d}$, 模型都会生成一个对应的情感极性标签 $y\in\{\text{positive},\text{neutral},\text{negative}\}$, 最终构成情感预测序列 $\boldsymbol{Y}$。

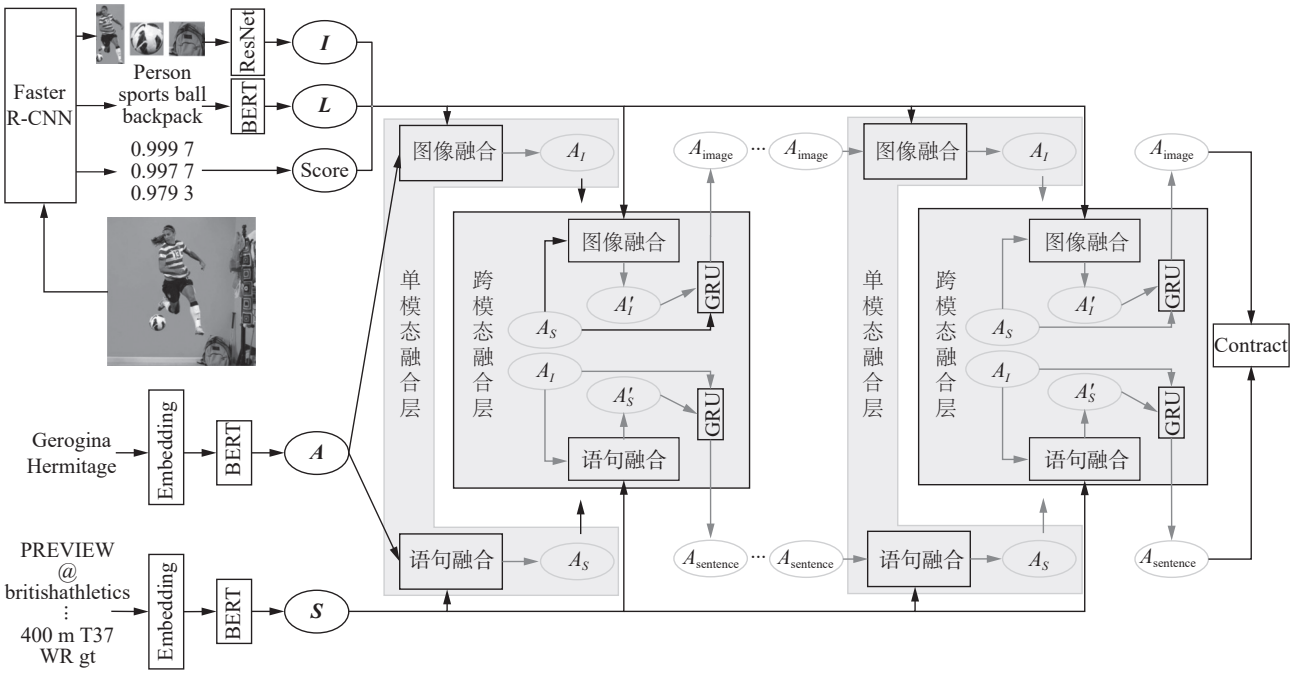

图 1　OAB-ALMSA 模型的整体结构

Fig. 1　Overall structure of OAB-ALMSA

2.1　嵌入层

2.1.1　文本信息

对于数据集中的每一个元素 $\boldsymbol{d}$, 都包含有一条语句 $\boldsymbol{S}$ 和一个方面词 $\boldsymbol{A}$。对于 $\boldsymbol{A}$, 本文通过可训练的 BERT 模型[24]来获取其向量化表示:

$$\boldsymbol{H}_A=\text{BERT}(\boldsymbol{A})\in\mathbf{R}^{l_1\times 768} \tag{1}$$

其中 l_1 为每个方面词被限定的长度。

而在对语句 $\boldsymbol{S}$ 的处理上, 本文参考了文献 [22] 中的做法, 将 $\boldsymbol{S}$ 分割为两部分, 分别为方面词 $\boldsymbol{A}$ 以及将语句中的方面词替换为单词“T”后所得到的新语句 $\boldsymbol{S}'=\{\omega_1,\omega_2,\cdots,\omega_a,\$\text{T}\$,\omega_{a+m+1},\cdots,\omega_o\}$, 之后将两者组合并添加标识符得到该样本最终作为输入的语句 $\boldsymbol{S}''=\{[\boldsymbol{CLS}],\boldsymbol{S}',[\boldsymbol{SEP}],\boldsymbol{A},[\boldsymbol{SEP}]\}$。之后, 与 $\boldsymbol{A}$ 的处理相似, 本文通过可训练的 BERT 模型对重新构建的语句 $\boldsymbol{S}''$ 进行向量化, 最终得到语句信息的向量化表示:

$$\boldsymbol{H}_S=\text{BERT}(\boldsymbol{S}'')\in\mathbf{R}^{l_2\times 768} \tag{2}$$

其中 l_2 为语句被限定的长度。

2.1.2　图像信息

对于数据集中的元素 $\boldsymbol{d}$ 来说, 其中所包含的每一条语句都会有一张与之对应的图像 $\boldsymbol{I}\in\mathbf{R}^{3\times H_0\times W_0}$, 本文通过使用目标检测算法 Faster R-CNN[25] 对该数据进行初步处理, 对于检测到的图像中的每一个目标, 可以获得与之对应的图像信息 $\boldsymbol{I}^i$、目标标签信息 $\boldsymbol{L}^i$、目标置信度信息 $\boldsymbol{S}^i_{\text{core}}$。以此为基础构建的图像整体信息, 包含所有目标的图像信息 $\boldsymbol{I}'=\{\boldsymbol{I}^0,\boldsymbol{I}^1,\cdots,\boldsymbol{I}^n\}$, 以及对应于每一个目标图像的标签信息 $\boldsymbol{L}'=\{\boldsymbol{L}^0,\boldsymbol{L}^1,\cdots,\boldsymbol{L}^n\}$, 对应于每一个标签的置信度信息 $\boldsymbol{S}'_{\text{core}}=\{\boldsymbol{S}^0_{\text{core}},\boldsymbol{S}^1_{\text{core}},\cdots,\boldsymbol{S}^n_{\text{core}}\}$。其中 $\boldsymbol{I}'$、$\boldsymbol{L}'$、S'_{core} 分别包含原图像的图像信息 $\boldsymbol{I}^0=\boldsymbol{I}$、标签信息 $\boldsymbol{L}^0=\boldsymbol{N/A}$、置信度 $\boldsymbol{S}^0_{\text{core}}=1$, 以及所有检测到目标的图像信息 $\boldsymbol{I}^i$、标签信息 $\boldsymbol{L}^i$、置信度信息 $\boldsymbol{S}^i_{\text{core}},1\leqslant i\leqslant n$。

考虑到深度 CNN 模型现已在许多图像处理任务中取得良好表现, 可以捕获对任务有益的特征信息, 因此本文采用残差网络 ResNet[26] 进行图像数据的特征提取。对于图像信息 $\boldsymbol{I}'$ 中的每一条信息 $\boldsymbol{I}^i$, 本文首先将该数据的大小调整为固定的 224×224, 以适应 ResNet 的输入要求。随后, 将调整后的数据输入到 ResNet 模型中, 并以最后一个

卷积层的输出作为图像信息的特征表示：

$$\boldsymbol{H}_I^i = \text{ResNet}(\boldsymbol{I}^i) \in \mathbf{R}^{49\times 2048} \tag{3}$$

通过该方法最终得到原始图像对应的图像特征信息 $\boldsymbol{H}_I = \{\boldsymbol{H}_I^0, \boldsymbol{H}_I^1, \cdots, \boldsymbol{H}_I^n\}$。

对于每一个图像的特征信息 $\boldsymbol{H}_I^i$ 所对应的标签信息 $\boldsymbol{L}^i$，本文采用与处理方面词 $\boldsymbol{A}$ 相似的方法，通过可训练的 BERT 模型对方面词进行词嵌入处理，得到每一个标签 $\boldsymbol{L}^i$ 所对应的向量化表示：

$$\boldsymbol{H}_L^i = \text{BERT}(\boldsymbol{L}^i) \in \mathbf{R}^{l_2\times 768} \tag{4}$$

其中每个标签被限定的长度与方面词相同，均为 l_2。通过该方法最终得到原始图像对应的标签特征信息 $\boldsymbol{H}_L = \{\boldsymbol{H}_L^0, \boldsymbol{H}_L^1, \cdots, \boldsymbol{H}_L^n\}$。

2.2 方面级的单模态融合层

2.2.1 目标注意力

为了充分利用目标检测算法所得到的原始图像中的细节信息，本文引入目标注意力机制来实现图像信息和文本信息之间的融合。目标注意力机制以基本的注意力机制[27]为基础构建，以 $\boldsymbol{H}_A$、$\boldsymbol{H}_I$、$\boldsymbol{S}'_{\text{core}}$ 为输入，以方面词和图像的融合结果 $\boldsymbol{A}_{\text{TT}}$ 为输出结果。$\boldsymbol{A}_{\text{TT}}$ 可表示为

$$\boldsymbol{A}_{\text{TT}} = \text{Object\_Attention}(\boldsymbol{H}_A, \boldsymbol{H}_I, \boldsymbol{H}_I, \boldsymbol{S}'_{\text{core}}) \tag{5}$$

该过程可以被具体地划分为 3 个部分。首先，通过权重矩阵对输入的特征向量信息进行初步的处理，得到

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{W}^q\boldsymbol{H}_A, \boldsymbol{W}^k\boldsymbol{H}_I, \boldsymbol{W}^v\boldsymbol{H}_I \tag{6}$$

式中：$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$ 为经过初始化的权重矩阵，在模型训练过程中进行参数优化。其次通过

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{K}^{\text{T}} \tag{7}$$

计算得到方面词对图像信息的注意力矩阵，其中 $x_{ij} \in \boldsymbol{X}$ 可以代表方面词中的第 i 个单词对第 j 个目标图像所表现出的注意力。然而，若将该矩阵直接用于计算中，则会丢失图像通过目标检测方法所得到的一部分先验知识，即图像中所包含的目标的置信度。一般而言，拥有更高置信度的目标所占据的图像区域包含的图像信息具有更高的清晰度以及更加明确的特征信息，因此应当获得更多的注意，即应当具有更大的注意力值。为此，本文在通过注意力机制来融合不同信息时，引入了目标检测方法所得到的置信度 $\boldsymbol{S}'_{\text{core}}$，以优化改进注意力矩阵的计算，通过

$$\boldsymbol{X}' = \boldsymbol{X} \cdot \boldsymbol{S}'_{\text{core}} = \boldsymbol{Q}\boldsymbol{K}^{\text{T}}\boldsymbol{S}'_{\text{core}} \tag{8}$$

计算得到新的注意力矩阵 $\boldsymbol{X}'$。最后一步是对注意力矩阵进行归一化处理，并利用注意力矩阵对图像特征向量进行加权求和，从而得到方面词和图像的融合结果：

$$\boldsymbol{A}_{\text{TT}} = \boldsymbol{X}'\boldsymbol{V} = \text{Soft}\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}\boldsymbol{S}_{\text{core}}}{\sqrt{l}}\right)\boldsymbol{V} \tag{9}$$

其中 l 为 $\boldsymbol{Q}$ 的维度。

2.2.2 图像融合单元

作为单模态融合层的一部分，该单元的主要功能是基于目标注意力机制完成方面词和图像信息的融合。为了达成这一目的，该单元首先利用上文介绍的目标注意力机制，将方面词与图像中所有目标的图像信息进行融合。得到：

$$\boldsymbol{A}_{I_1} = \text{Object\_Attention}(\boldsymbol{H}_A, \boldsymbol{H}_I, \boldsymbol{H}_I, \boldsymbol{S}'_{\text{core}}) \tag{10}$$

此外，考虑到目标标签信息 $\boldsymbol{L}'$ 是对图像中检测出的目标进行的文字标注，是对目标的分类，它概括了图像中目标的主要特征，因此也可以在一定程度上代表该目标所对应的图像信息。该标签信息与图像本身的图像信息相比，所包含的信息较为简单，但相对来说也更加清晰，易于理解。因此，该模块通过将标签信息引入到目标注意力机制中，计算得到：

$$\boldsymbol{A}_{I_2} = \text{Object\_Attention}(\boldsymbol{H}_A, \boldsymbol{H}_L, \boldsymbol{H}_I, \boldsymbol{S}'_{\text{core}}) \tag{11}$$

为了统一上述两种融合结果，该单元选择使用门控循环单元 (gate recurrent unit, GRU) 网络[28]来对两者进行融合操作，得到最终的方面词和图像信息的融合结果：

$$\boldsymbol{A}_I = \text{GRU}(\boldsymbol{A}_{I_1}, \boldsymbol{A}_{I_2}) \tag{12}$$

2.2.3 语句融合单元

作为单模态融合层的一部分，该单元的主要功能是基于注意力机制完成方面词和语句信息的融合。通过

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{W}^q\boldsymbol{H}_A, \boldsymbol{W}^k\boldsymbol{H}_S, \boldsymbol{W}^v\boldsymbol{H}_S \tag{13}$$

对输入数据进行加权化处理。之后，计算得到方面词所对应的语句的综合信息 $\boldsymbol{A}_S$，将 $\boldsymbol{A}_S$ 作为方面词和语句的融合结果。

$$\boldsymbol{A}_S = \text{Attention}(\boldsymbol{H}_A, \boldsymbol{H}_S, H_S) = \text{Soft}\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{l}}\right)\boldsymbol{V} \tag{14}$$

2.3 跨模态融合层

方面级的单模态融合层所得到的结果 $\boldsymbol{A}_I$ 和 $\boldsymbol{A}_S$ 分别为方面词和语句、方面词和图像之间的信息融合结果。因此，该层主要以跨模态融合为目的完成方面词、语句、图像信息之间的融合。在具体实现上，仍以注意力机制和目标注意力机制为核心，这一过程主要分为两个部分。首先参考式 (5)、(9)、(14)，我们完成方面词和图像信息的融合结果 $\boldsymbol{A}_I$ 和语句特征信息 $\boldsymbol{H}_S$ 的融合，方面词和语句的融合结果 $\boldsymbol{A}_S$ 和图像特征信息 $\boldsymbol{H}_I$、$\boldsymbol{H}_L$、$\boldsymbol{S}'_{\text{core}}$ 之间的融合：

$$\boldsymbol{A}'_S = \text{Attention}(\boldsymbol{A}_I, \boldsymbol{H}_S, \boldsymbol{H}_S) \tag{15}$$

$$\boldsymbol{A}'_{I_1} = \text{Object\_Attention}(\boldsymbol{A}_S, \boldsymbol{H}_I, \boldsymbol{H}_I, \boldsymbol{S}'_{\text{core}}) \tag{16}$$

$$\boldsymbol{A}'_{I_2} = \text{Object\_Attention}(\boldsymbol{A}_S, \boldsymbol{H}_L, \boldsymbol{H}_I, \boldsymbol{S}'_{\text{core}}) \tag{17}$$

$$\boldsymbol{A}'_I = \text{GRU}(\boldsymbol{A}'_{I_1}, \boldsymbol{A}'_{I_2}) \tag{18}$$

在一定程度上, $\boldsymbol{A}'_I$和$\boldsymbol{A}'_S$都可以被认为同时包含方面词的特征信息、语句的特征信息、图像的特征信息, 即均可以代表方面词、语句、图像三者的信息融合结果, 但两者本质上都只是图像特征信息或语句特征信息。因此, 本文通过 GRU 网络计算得到的融合结果:

$$\boldsymbol{A}_{\text{image}} = \text{GRU}(\boldsymbol{A}_S, \boldsymbol{A}'_I) \tag{19}$$

$$\boldsymbol{A}_{\text{sentence}} = \text{GRU}(\boldsymbol{A}_I, \boldsymbol{A}'_S) \tag{20}$$

作为方面级的多模态信息融合结果。其中 $\boldsymbol{A}_{\text{image}}$ 和 $\boldsymbol{A}_{\text{sentence}}$ 均代表了方面词、语句和图像充分融合的结果。

2.4　迭代层

上述操作在功能上已经达到了方面级的多模态信息融合的目的, 但尚不足以充分提取出隐藏于原始数据中的特征信息。因此, 本文通过构建迭代结构, 对方面级单模态融合层和跨模态融合层所组合而成的信息融合层进行多次使用和连接, 以提高方面级的多模态融合信息的有效性。

为使原本的信息融合层能够完整地构成迭代结构, 本文对每一层的输入输出进行了统一化的定义。对于第一个融合层来说, 其主要输入为方面词的特征信息, 而输出则有两种形式, 均用于代表方面级多模态融合结果的 $\boldsymbol{A}_{\text{image}}$、$\boldsymbol{A}_{\text{sentence}}$ 特征信息。两者均是多种模态信息融合所得到的结果, 可以有效地筛选出原始图像和语句特征信息中的有效部分, 即可以起到与输入的方面词特征信息相同的作用。因此本文通过将式 (10)、(11)、(14) 中的 $\boldsymbol{H}_A$ 在形式上替换为 $\boldsymbol{A}_{\text{image}}$ 和 $\boldsymbol{A}_{\text{sentence}}$ 的方式, 以实现各层结构之间形式上的统一。

迭代结构的不同层之间通过 $\boldsymbol{A}_{\text{image}}$ 和 $\boldsymbol{A}_{\text{sentence}}$ 来完成信息的交流, 第一层输入的 $\boldsymbol{A}_{\text{image}}$ 和 $\boldsymbol{A}_{\text{sentence}}$ 在内容上等于 $\boldsymbol{H}_A$。

2.5　情感预测层

将最后一个融合层所得到的信息 $\boldsymbol{A}_{\text{image}}$ 和 $\boldsymbol{A}_{\text{sentence}}$ 拼接后送到线性层中, 用于情感预测, 取输出中概率最高的标签作为最终结果:

$$\hat{y} = \text{Linear}(\text{contract}(\boldsymbol{A}_{\text{image}}, \boldsymbol{A}_{\text{sentence}})) \tag{21}$$

2.6　模型训练

为了优化本文所述模型中所有的参数, 本文的目标是最小化交叉熵损失函数:

$$L_{\text{oss}} = -\sum_{i=1}^{3} y^i \log(\hat{y}^i) \tag{22}$$

式中: $\hat{y}$ 为模型最终预测得到的情感标签, y 为实际的情感标签。

2.7　课程式学习

课程式学习 CL(curriculum learning)主要包含两部分[29]: 难易度评估(difficulty evaluation)和课程设计(curriculum arrangement)。对于给定的数据集 $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n\}$, 难易度评估的目标是为数据集中的每一个元素 $\boldsymbol{d}_j$ 计算出一个对该元素的评价值 $\boldsymbol{e}_j$, 用于反映该数据在使用模型进行分析时所体现出的难易度, 以此构成该数据集 $\boldsymbol{D}$ 的难易度评估集合 $\boldsymbol{E}$。课程设计以难易度评估集合 $\boldsymbol{E}$ 为基础, 将数据集划分成多个部分并按照一定的顺序构成数据集 $\boldsymbol{D}$ 的多个子集 $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_n\}$, 并按照集合 $\boldsymbol{C}$ 的顺序对模型进行训练。

2.7.1　难易度评估

在对样本难易度进行评估时不仅要考虑样本本身的复杂度, 还需要样本在使用模型进行分析时所表现出的难易度。因此, 本文先对模型 OAB-ALMSA 进行预训练, 对于预训练后的最终模型, 我们取其嵌入层的部分模型结构和参数来完成方面词、语句和图像中目标标签信息的特征提取, 并获取其对应的向量化表示, 进而计算样本数据的难易度。

方面词、语句和图像之间的复杂关系是导致方面级多模态情感分析任务困难性的主要原因之一, 因此本文通过计算三者之间的相关性对数据分析的难易度进行评估。考虑到图像的复杂性是导致方面词、语句和图像之间复杂关系的一大原因, 因此本文通过计算分别得到方面词和图像特征信息、语句和图像特征信息之间的相关性, 以此作为样本数据复杂度的评价标准。在具体计算过程中, 本文选择使用图像中通过目标检测算法提取的目标特征信息代替图像的整体特征信息。同时, 考虑到图像中目标的标签信息与方面词和语句具有同样的文本表现形式, 并且与目标的图像信息相比内容更加明确, 因此本文选择使用目标标签特征信息来代表目标的图像特征信息, 进而与方面词和语句特征信息进行相关性计算。通过计算得到方面词与图像之间的相关性 $\boldsymbol{e}_1$, 语句与图像之间的相关性 $\boldsymbol{e}_2$:

$$\boldsymbol{e}_1 = \frac{\boldsymbol{H}_S \cdot \boldsymbol{H}_L}{\|\boldsymbol{H}_S\| * \|\boldsymbol{H}_L\|} \tag{23}$$

$$\boldsymbol{e}_2 = \frac{\boldsymbol{H}_A \cdot \boldsymbol{H}_L}{\|\boldsymbol{H}_A\| * \|\boldsymbol{H}_L\|} \tag{24}$$

2.7.2　课程设计

课程设计的第一步便是通过得到的难易度评估值对数据集进行切分，考虑到上文的定义中用于评价样本难易度的指标有$\boldsymbol{e}_1$和$\boldsymbol{e}_2$两种，因此本文在对数据集划分时选择以$\boldsymbol{e}_1$和$\boldsymbol{e}_2$为分类指标的同时以$\boldsymbol{e}_1+\boldsymbol{e}_2$为辅助对分类方案加以选择。

对于整个数据集$\boldsymbol{D}$计算所得的评价指标集$\boldsymbol{E}=\{(\boldsymbol{e}_1^1,\boldsymbol{e}_2^1),(\boldsymbol{e}_1^2,\boldsymbol{e}_2^2),\cdots,(\boldsymbol{e}_1^n,\boldsymbol{e}_2^n)\}$，本文选择使用 K-means 聚类的方式将原数据集$\boldsymbol{D}$划分为k个集合。之后，对于得到的每一个集合，计算样本的评估值$\boldsymbol{e}_1$和$\boldsymbol{e}_2$的平均值$\bar{\boldsymbol{e}}_1$、$\bar{\boldsymbol{e}}_2$，并以$\bar{\boldsymbol{e}}$为指标将集合以从大到小的顺序进行排序，得到$\boldsymbol{D}=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_k\}$。

$$\bar{\boldsymbol{e}}=\bar{\boldsymbol{e}}_1+\bar{\boldsymbol{e}}_2 \tag{25}$$

此时可以认为集合的序号越小，该集合的样本越容易进行本文模型的情感分析。在理想状态下，这些集合应当满足：

$$e_1^s+e_2^s\geqslant e_1^t+e_2^t|\forall d^s\in D_i,d^t\in D_j,i<j \tag{26}$$

若存在s、t不满足上述条件，则可以认为这两个样本的划分存在错误，本文通过统计整个集合$\boldsymbol{D}$中分类存在错误的样本数量来计算得到平均每个子集$\boldsymbol{D}_i$中存在的分类错误的样本数量，对 K-means 分类结果进行评价，得到的结果如图 2 所示。

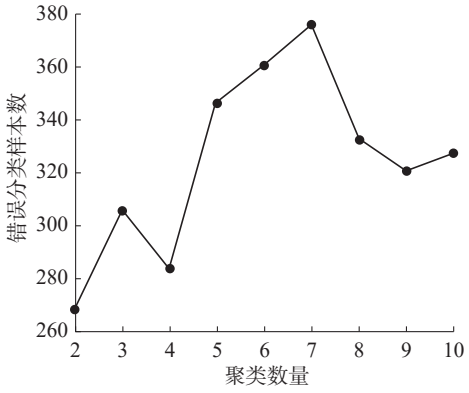

图 2　聚类数量对平均错误样本数量的影响

Fig. 2　Impact of cluster quantity on the average number of incorrect samples

从该折线图可知，当聚类数量k达到 5 时，分类错误的数量出现急速增加。因此，本文在进行实验时，选择将 K-means 聚类的k值定为 2、3、4，并在比较不同的k值对最终结果所带来的影响后最终将k值定为 4。

确定k值后，便可以通过 K-means 聚类算法以$\boldsymbol{e}_1$和$\boldsymbol{e}_2$为指标将集划分为$\boldsymbol{D}=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_k\}$，并构建$\boldsymbol{D}$的子集集合$\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$，其中

$$\boldsymbol{C}_i=\bigcup_{j=1}^{i}\boldsymbol{D}_j \tag{27}$$

然后根据小步算法（Baby Step），首先以$\boldsymbol{C}_1$为训练集对 OAB-ALMSA 模型进行训练，在完成p轮训练之后，以$\boldsymbol{C}_2$为训练集继续进行训练，以此类推，直到完成$\boldsymbol{C}_n$的训练，并在最后完整的训练集$\boldsymbol{C}_n$上进行多次额外训练。

3　实验与讨论

3.1　数据集和实验设置

实验主要在公开的基准数据集 TWITTER-2015[22] 上进行。Twitter 因其本身具有的多模态信息、高度情绪化的信息以及与现实事件高度相关的特点，成为方面级的多模态情感分析任务理想的数据来源。该数据集主要包含 2014—2015 年用户所发布的 Twitter 帖子[30]，其中涉及的方面词分属 4 个类别：人、地点、组织和其他。每一条数据都包含语句以及与之对应的图像，并标注了决定情感分析方向的方面词以及图文信息在该方面所表现出的情感倾向。该数据集中所包含的训练集、验证集、测试集的统计数据如表 1 所示。在模型训练过程中，我们将学习率初始化为3×10^{-5}，并在训练过程中对其进行优化调整，批处理大小设置为 10，对于模型中的课程式学习部分来说，总迭代次数为 12，小步算法中每一小步的迭代次数为 2。模型使用 BERTAdam 优化器来完成模型参数的更新，整个模型采用 PyThon 语言和 PyTorch 框架完成。在对模型性能的评估上，本文选择使用准确率 ACC(accuracy) 和 F_1 分数作为评价标准。

表 1　实验数据统计

Table 1　Experimental data statistics

数据集	积极	中立	消极	总计
训练集	928	1883	368	3 179
验证集	303	670	139	1 122
测试集	317	607	113	1 037

3.2　对比实验

本节将 OAB-ALMSA 模型与现有的 ALMSA 模型中较为有代表性的几种方法进行了比较，表 2 给出了每一种方法的准确率和 F_1 分数。首先对几种仅考虑了图像或语句的方法进行了比较。

1）Res-Aspect-ATT，本文所构建的模型仅处理方面词信息和图像信息，分别通过 ResNet 和 BERT 对图像和方面词信息进行特征提取，并将提取出的信息通过一个注意力层进行特征融合。

2）TD-LSTM (temporal dependence-based long short-term memory network)[30]，通过两个 LSTM 模型分别对语句中位于方面词左侧和右侧的信息进行特征提取，并通过连接最后一个隐藏层的向量

来完成情感分析。

3）BERT-Aspect-ATT (BERT-aspect-attention), 本文所构建的模型仅处理方面词信息和语句信息, 通过 BERT 完成方面词和语句的词嵌入, 并将得到的特征信息通过单注意力层来完成特征融合。

4）MIMN (multi-interactive memory network)[1], 通过注意力机制和两组交互记忆网络完成方面词和语句、图像之间的融合。

5）TomBERT (target-oriented multimodal BERT)[22], 该模型整体上由 3 个 BERT 模块组成, 这些 BERT 模块分别被用于获取方面词和图像的交互结果, 提取语句的特征信息, 完成多模态之间的信息融合。

6）ModalNet-BERT (multimodal fusion discriminant attentional network)[3], 基于注意力机制来完成单模态的方面级情感分析, 之后通过构建融合判别矩阵对不同模态的融合结果进行进一步的融合。

7）EF-CapTrBERT (early fusion caption Transformer BERT)[4], 基于 Transformer 将图像转化为辅助语句, 通过 BERT 模型实现方面词、辅助语句、原始语句之间的融合。

8）EF-CapTrBERT-DE (early fusion caption Transformer BERT domain-specific)[4], 与 EF-CapTrBERT 相比, 选择使用文献 [31] 中特定域的权重初始化 BERT, 但是不改变 Transformer 部分的数据。

3.2.1　对比实验主要结果

模型 OAB-ALMSA、引入课程式学习后的模型 OAB-ALMSA-CL 以及多种基准模型的实验结果均如表 2 所示, 通过分析表 2, 可以发现本文模型 OAB-ALMSA-CL 取得了目前最高的 F_1, 这表明该模型在 TWITTER-2015 数据集上所得到的预测结果在综合考虑了精准率和召回率时表现最佳。

表 2　对比实验结果

Table 2　Results of contrast experiment

模态	模型	ACC	Mac-F_1
图像	Res-Aspect-ATT	0.5265	0.3322
文本	TD-LSTM	0.6830	0.6143
	BERT-Aspect-ATT	0.7705	0.7180
图像+文本	MIMN	0.7353	0.6649
	TomBERT	0.7715	0.7180
	ModalNet-BERT	0.7903	0.7250
	EF-CapTrBERT	0.7801	0.7325
	EF-CapTrBERT-DE	0.7792	0.7390
	OAB-ALMSA	0.7839	0.7427
	OAB-ALMSA-CL	0.7869	**0.7479**

对于所给定的基准模型来说, 仅考虑图像信息的 Res-Aspect-ATT 模型有着很低的准确率和 F_1, 这表明仅从图像中提取的特征信息本身并不适合用于情感分析。与 Res-Aspect-ATT 模型相比, 仅考虑了语句信息的 TD-LSTM 模型在准确率和 F_1 上均有 10%~30% 的提升, 这在一定程度上证明了在 TWITTER-2015 数据集上进行情感分析任务时语句所带有的信息的重要性。对于引入了 BERT 作为文本特征提取工具和注意力机制来完成方面词和语句的信息融合的 BERT-Aspect-ATT 模型来说, 其在结果的准确率和 F_1上都有显著提升, 这在很大程度上证明了 BERT 模型在文本特征提取以及注意力机制在进行信息融合时的有效性。

MIMN 作为最早用于处理方面级多模态情感分析的一批模型之一, 在 TWITTER-2015 数据集上的表现要比 BERT-Aspect-ATT 模型略差, 这表明在多模态情感分析过程中对图像信息不充分的处理可能会在一定程度上干扰模型从文本中提取信息, 从而造成模型对样本情感极性的错误判断, 进而影响模型效果。相较而言, 研究人员在此之后针对此问题所提出的各种模型都有更优秀的性能, 其中 ModalNet-BERT 模型通过构建融合判别矩阵来监督多模态融合过程, 在准确率上表现出色, 但并没有取得较好的 F_1。相比之下, 模型 OAB-ALMSA-CL 虽然在准确率上略低, 但 F_1 有了明显的提升, 高达 2.99%, 这表明与只使用注意力机制的模型 ModalNet-BERT 相比, 通过引入目标注意力机制以提高对原始图像中信息利用的模型 OAB-ALMSA-CL 能够显著地提高模型在数据集上准确率和召回率的平衡。与模型 ModalNet-BERT 相比, 模型 EF-CapTrBERT-DE 虽然在准确率上的表现并不突出, 但是却拥有当前最高的 F_1。而本文模型 OAB-ALMSA-CL 在与之的对比中, 分别在准确率和 F_1 上提高了 0.77% 和 0.89%, 这证明了本文所提出的目标注意力机制、模型结构以及针对性的课程式学习策略在方面级的多模态情感分析任务中的有效性。

3.2.2　迭代层数对模型性能的影响

对于本文模型 OAB-ALMSA 来说, 模型的迭代层数在很大程度上影响了模型的整体性能, 如图 3 所示。

随着层数的增加, 模型的准确率和 F_1 都呈现出先提升后下降最终趋于平稳的趋势。模型性能的上升阶段集中于层数在 3 及以下的情况; 当层数达到 4 时, 模型的准确率保持相对稳定, 而 F_1 出现了明显下降; 而当层数达到 5 时, 模型的

准确率和 F_1 都有了明显的下降。当模型的层数达到 5 以上时，结果的准确率和 F_1 均趋于平稳，没有较大的变化。这表明，当迭代层数较少时，随着层数的增加，多模态信息之间的融合会越来越充分，最终的情感预测效果也会越来越好。但当迭代层数超过一定限度时，模型复杂度增加，容易出现过拟合现象，进而影响最终情感预测的结果。因此，在本文实验过程中，模型的迭代层数均不大于 4。

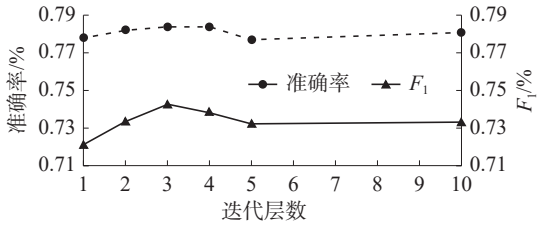

图 3　迭代层数对模型性能的影响

Fig. 3　Impact of iteration levels on model performance

3.3　消融实验

表 3 为本文的消融实验结果，所有消融实验均基于 OAB-ALMSA 模型进行。可以看出，模型 OAB-ALMSA 中的每一块结构对于最终效果都是不可或缺的。具体而言，当我们从 OAB-ALMSA 模型中删除 2.2.1 节引入的目标注意力机制，并将其替换为基础的注意力机制时，模型性能较 OAB-ALMSA 模型有着显著的降低，这证明了整个模型中最重要、对结果影响最大的结构是用于完成方面词和图像信息融合的目标注意力机制，这表明了从图像中提取目标的细节信息对于提取多模态信息中的情感特征至关重要，以及本文所提出的目标注意力机制在该方面能够起到有效的作用。对于每一层的迭代结构，无论是去除其中的方面级单模态融合层还是跨模态融合层，最终模型性能都有明显下降，在这种情况下，仅进行单模态与方面词的融合，或者直接进行多模态与方面词之间的融合都无法充分地提取多种模态信息之间有益于情感分析的特征。

表 3　消融实验结果

Table 3　Ablation experiment results

消融模型	ACC	Mac-F_1
OAB-ALMSA-CL	**0.7869**	**0.7479**
OAB-ALMSA	0.7839	0.7427
w/o 单模态融合层	0.7772	0.7259
w/o 跨模态融合层	0.7733	0.7222
w/o 迭代层	0.7782	0.7216
w/o Object-Attention	0.7676	0.7208

此外，通过对比 OAB-ALMSA 与 OAB-ALMSA-CL 模型的结果可以发现，本文所构建的课程式学习策略显著地提高了情感分析的性能。方面级的多模态情感分析是一个较为困难的任务，其训练所用样本数据具有的较高复杂度和差异性增加了训练的难度。合适的课程式学习策略使得模型可以先在较为简单的样本上进行学习，这减少了样本差异性对模型优化造成的困难，提高了最终模型的学习结果，即情感分析的性能。

3.4　案例分析

为了更好地了解模型 OAB-ALMSA 的优势，本文还将数据样本根据方面词的长度 lenAspect 对数据集进行了划分，得到方面词为单个词的数据集以及方面词由多个词构成的数据集，并统计了模型 OAB-ALMSA 以及该模型的部分消融模型所得到的结果在两种不同数据集上的表现，最终的结果如表 4 所示。

表 4　针对方面词长度的性能细分

Table 4　Performance segmentation for length of aspect

消融模型	lenAspect=1		lenAspect≥2	
	ACC	F_1	ACC	F_1
w/o Object-Attention	0.755	0.710	0.778	0.728
w/o 跨模态融合层	0.755	0.702	0.788	0.740
w/o 单模态融合层	0.744	0.689	0.788	0.740
OAB-ALMSA	0.772	0.733	0.789	0.745

根据表 4 中数据可以发现，模型 OAB-ALMSA 在长方面词的数据集上所表现出的优势并不明显，但在由短长度方面词组成的数据集上，其却明显优于消融模型。当方面词的长度仅为 1 时，其所包含的信息量要明显少于长度不小于 2 的方面词，从中获取信息来完成多模态信息融合的难度也就更高。而模型 OAB-ALMSA 通过对图像细节信息的利用提高了方面词和图像信息的融合效果，实验结果也表明该模型在一定程度上缓解了短方面词对方面级多模态情感分析任务所带来的负面影响。

图 4 列出了 OAB-ALMSA 模型及其部分消融实验对具体案例样本的预测结果之间的比较。对于图 4(a) 的案例 1，消融模型因其结构的不完整，不能很好地融合多模态信息，所以预测结果更偏向于纯文本信息所表现出的中立情感。然而 OAB-ALMSA 模型在提取了图像中与人名“Martin”相关的人物表情后，结合文本信息便分析出了其中所包含的消极信息，做出了正确的预测。

[Eddie] and the Faithfull Pearl Jam fans in Buenos Aires . Photo by @ epozzoni # PJSA2013
Human Label: (positive)

w/o 迭代层: (neutral ×)
w/o Object-Attention : (neutral ×)
OAB-ALMSA : (positive √)

(a) 案例1

OFFICIAL: [Martin] St . Louis announces retirement from the National Hockey League . # NYR

Human Label: (negative)
w/o 迭代层: (neutral ×)
w/o Object-Attention: (neutral ×)
OAB-ALMSA: (negative √)

(b) 案例2

图 4　实验案例

Fig. 4　Experimental examples

4　结束语

本文提出了一种基于目标注意力的方面级多模态情感分析模型, 通过引入目标检测算法提取了原始图像中目标细节信息, 并结合目标检测结果的特点对传统注意力机制进行了改进, 构建了目标注意力机制以完成图像细节信息与语句、方面词之间的融合。以此为基础, 本文构建了方面级的单模态融合层、跨模态融合层以及迭代结构来完成多模态信息在方面级上的充分融合, 然后得到最终的情感分类。此外, 本文还针对训练样本的复杂性问题提出了合适的课程式学习策略, 提升了模型在方面级多模态情感分析任务中的性能。

由于本文使用的基准数据集所含有的方面词大多为固有名词, 因此限制了方面词与图像目标信息的融合效果。下一步的工作将探索对方面词的预处理方法来提高方面词的有效应用。本文目前所使用到的模型参数较为复杂, 训练效率不高, 在下一步的工作中也将对此进行改进。

参考文献:

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[J]. 国家图书馆学刊, 2023, 32(2): 39.

[2] XU Nan, MAO Wenji, CHEN Guandan. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI Press, 2019: 371−378.

[3] ZHANG Zhe, WANG Zhu, LI Xiaona, et al. ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network[J]. World wide web, 2021, 24(6): 1957−1974.

[4] KHAN Z, FU Yun. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event China: ACM, 2021: 3034−3042.

[5] 王伟贤, 吴俊. 基于情感词典与语义规则集的微博文本情感分析[J]. 计算机科学与应用, 2023, 13(4): 754−763.
WANG Weixian, WU Jun. Sentiment analysis of microblog text based on sentiment dictionary and semantic rule set[J]. Computer science and application, 2023, 13(4): 754−763.

[6] SHANG Yongmin, ZHAO Yuqin. Sentiment analysis and implementation of online reviews based on machine learning[J]. Journal of Dali University, 2021, 6(12): 80−86.

[7] HUANG Jiaxin, MENG Yu, GUO Fang, et al. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020.

[8] KIRITCHENKO S, ZHU X, CHERRY C, et al. Detecting aspects and sentiment in customer reviews[C]//8th International Workshop on Semantic Evaluation. [S.l.]: [s.n.], 2014, 437−442.

[9] 王婷, 杨文忠. 文本情感分析方法研究综述[J]. 计算机工程与应用, 2021, 57(12): 11−24.
WANG Ting, YANG Wenzhong. Review of text sentiment analysis methods[J]. Computer engineering and applications, 2021, 57(12): 11−24.

[10] LONDHE A, RAO P V R D P. Aspect based sentiment analysis–an incremental model learning approach using LSTM-RNN[M]//Communications in Computer and Information Science. Cham: Springer International Publishing, 2021: 677−689.

[11] XUE Wei, LI Tao. Aspect based sentiment analysis with gated convolutional networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 2514−2523.

[12] 张文轩, 殷雁君, 智敏. 用于方面级情感分析的情感增强双图卷积网络[J]. 计算机科学与探索, 2024, 18(1): 217−230.
ZHANG Wenxuan, YIN Yanjun, ZHI Min. Affection enhanced dual graph convolution network for aspect based sentiment analysis[J]. Journal of frontiers of computer science and technology, 2024, 18(1): 217−230.

[13] WANG Bailin, LU Wei. Learning latent opinions for aspect-level sentiment classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 5537−5544.

[14] 曾锋, 曾碧卿, 韩旭丽, 等. 基于双层注意力循环神经网络的方面级情感分析[J]. 中文信息学报, 2019, 33(6): 108−115.

ZENG Feng, ZENG Biqing, HAN Xuli, et al. Double attention neural network for aspect-based sentiment analysis[J]. Journal of Chinese information processing, 2019, 33(6): 108−115.

[15] 谢珺, 王雨竹, 陈波, 等. 基于双指导注意力网络的属性情感分析模型[J]. 计算机研究与发展, 2022, 59(12): 2831−2843.
XIE Jun, WANG Yuzhu, CHEN Bo, et al. Aspect-based sentiment analysis model with Bi-guide attention network[J]. Journal of computer research and development, 2022, 59(12): 2831−2843.

[16] 李丽双, 周安桥, 刘阳, 等. 基于动态注意力 GRU 的特定目标情感分类[J]. 中国科学 (信息科学), 2019, 49(8): 1019−1030.
LI Lishuang, ZHOU Anqiao, LIU Yang, et al. Aspect-based sentiment analysis based on dynamic attention GRU[J]. Scientia sinica (informationis), 2019, 49(8): 1019−1030.

[17] WU Zhengxuan, ONG D C. Context-guided BERT for targeted aspect-based sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2021: 14094−14102.

[18] 曾凡旭, 李旭, 姚春龙, 等. 基于 BERT 的端到端方面级情感分析[J]. 大连工业大学学报, 2022, 41(3): 228−234.
ZENG Fanxu, LI Xu, YAO Chunlong, et al. End-to-end aspect-level sentiment analysis based on BERT[J]. Journal of Dalian Polytechnic University, 2022, 41(3): 228−234.

[19] GU Donghong, WANG Jiaqian, CAI Shaohua, et al. Targeted aspect-based multimodal sentiment analysis: an attention capsule extraction and multi-head fusion network [J]. IEEE access, 2021, 9: 157329−157336.

[20] 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.

[21] ZHOU Jie, ZHAO Jiabao, HUANG J X, et al. MASAD: a large-scale dataset for multimodal aspect-based sentiment analysis[J]. Neurocomputing, 2021, 455: 47−58.

[22] YU Jianfei, JIANG Jing. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization, 2019: 5408−5414.

[23] LING Yan, YU Jianfei, XIA Rui. Vision-language pre-training for multimodal aspect-based sentiment analysis [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 2149−2159.

[24] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018−11−11)[2021−01−01]. http://arxiv.org/abs/1810.04805.

[25] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137−1149.

[26] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017−06−12)[2021−01−01]. http://arxiv.org/abs/1706.03762.

[28] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014−12−11)[2021−01−01]. http://arxiv.org/abs/1412.3555.

[29] WANG Xin, CHEN Yudong, ZHU Wenwu. A survey on curriculum learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(9): 4555−4576.

[30] ZHANG Qi, FU Jinlan, LIU Xiaoyu, et al. Adaptive co-attention network for named entity recognition in tweets[J]. Proceedings of the AAAI conference on artificial intelligence. New Orleans: AAAI Press, 2018: 5674−5681.

[31] NGUYEN D Q, VU T, TUAN NGUYEN A. BERTweet: a pre-trained language model for English Tweets[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2020: 9−14.

作者简介：

朱超杰，硕士研究生，主要研究方向为多模态情感分析。E-mail: zcj1191098231@163.com。

闫昱名，北京华电电子商务科技有限公司工程师，主要研究方向为信息化项目管理及实施。E-mail: yanyuming@chd.com.cn。

高小燕，讲师，博士，主要研究方向为自然语言处理。E-mail: gaoxiaoyan@bjut.edu.cn