基于方面引导的图文渐进融合的多模态方面级情感分析方法

闫自达[1,2]，郭军军[*1,2]，余正涛[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
2641668065@qq.com, guojjgb@163.com
ztyu@hotmail.com

摘要

多模态方面级情感分析旨在通过结合图像信息和文本信息来识别特定方面的情感极性。然而，图像和文本作为两种不同的模态，其在数据表现形式和语义表达上存在显著差异，缩小模态鸿沟和跨模态特征融合是多模态方面级情感分析任务中出现的两个关键问题。对此，本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，该方法采用图像和文本中重叠的方面信息作为枢轴，利用方面引导的图文对比学习和基于对比的跨模态语义交互来缩小模态差异、促进语义交互，然后在多模态特征空间中整合视觉和文本信息，通过方面引导的基于对比的多模态语义融合来促进跨模态特征融合，从而提升多模态情感分析的性能。在三个多模态方面级情感分析基准数据集上的实验结果证明了本文提出方法的有效性，并且优于其他大多数最先进的多模态方面级情感分析方法。

关键词： 情感分析；语义交互；对比学习；多模态融合

Aspect-Guided Progressive Fusion of Text and Image for Multimodal Aspect-Based Sentiment Analysis

Zida Yan[1,2], Junjun Guo[*1,2], Zhengtao Yu[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
2641668065@qq.com, guojjgb@163.com
ztyu@hotmail.com

Abstract

Multimodal Aspect-Based Sentiment Analysis aims to identify the sentiment polarity of specific aspects by combining image and text information. However, images and text are two types modalities, exhibit significant differences in data representation and semantic expression. Narrowing the modal gap and cross-modal feature fusion are two key challenges in multimodal aspect-based sentiment analysis. To address these issues, this paper proposes a aspect-guided progressive fusion of text and image for multimodal aspect-based sentiment analysis method, which uses the overlapping aspect information in visual and textual modalities as a pivot, utilizes aspect-guided contrastive cross-modal semantic interaction to narrow the modal differences, and then integrates visual and textual information in the multimodal feature space. It promotes cross-modal feature fusion through aspect-guided contrastive multimodal semantic fusion, thereby

*郭军军（通讯作者）：guojjgb@163.com
基金项目：国家自然科学基金（62162037, 62266027, U21B2027, 62266028）；云南省基础研究计划项目（202301AT070444,202001AT070047, 202001AT070046）

enhancing the performance of multimodal sentiment analysis. Experimental results on three multimodal aspect-based sentiment analysis benchmark datasets have proven the effectiveness of the proposed method, which outperforms most other state-of-the-art multimodal aspect-based sentiment analysis methods.

Keywords: Sentiment analysis , Semantic interaction , Contrastive learning , Multimodal fusion

1 引言

多模态方面级情感分析（*Multi-modal Aspect-Based Sentiment Analysis*, MABSA）的目标是将视觉信息融入文本中，以准确地识别句子中特定方面的情感极性。这一领域近年来获得了国内外研究者的广泛关注，相较于传统的仅基于文本的方面级情感分析，通过整合图像信息，能显著提升情感分析的性能。这项技术在社交媒体、医疗保健、教育等多个领域均展现出广泛的应用前景。

在多模态方面级情感分析研究中，图文两种模态间存在明显的语义差距，跨模态的表示学习和语义对齐通常较难。例如，如图1所示，文本描述了结婚的事件，而图像则展示了一位悲伤的男子。文本和图像之间存在显著的语义鸿沟，只有少数重叠的语义枢纽，如实体“Jennifer”和“Ross”以及视觉表达如“cryin”，才能用来弥合这两种模态之间的语义鸿沟。

图 1. 多模态方面级情感分析任务的示例。给定一张图片、一条推文和特定方面目标，我们的目标旨在预测每个方面的情感极性。

最近的许多研究通过采用几种面向方面的多模态融合策略来应对MABSA中的视觉-文本的对齐挑战，如基于方面的跨模态注意力 (Yu and Jiang, 2019; Yu et al., 2020; Zhang et al., 2021)、视觉-文本相关的多任务学习(Yu et al., 2022b)、视觉-文本跨模态预训练(Yu et al., 2022a)等。然而，这些方法中的大多数倾向于面向方面的特征级融合，直接将视觉、文本和方面表示作为输入，通常忽略了不同模态之间的表示差距。

跨模态语义交互和多模态特征融合是多模态融合过程中必须考虑的两个相互关联的问题。通过在模态空间中进行跨模态语义交互对齐，可以更容易地在多模态特征空间中融合它们的表示。对齐和融合是两个相互依存的子任务，对于成功的多模态融合都是必不可少的。此外，方面通常是图像和文本的重叠语义信息，可作为促进模态语义交互和特征融合的合适枢纽。

为了解决MABSA任务中出现的模态对齐难融合难的问题，本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，采用方面为枢纽，引入基于方面的三层对比学习，旨在逐层细化模态对齐，以层次化地促进图文特征融合。提出方面引导的图文对比学习来缩小模态差异；提出方面引导的基于对比的跨模态语义交互模块，通过双层跨模态注意力机制来促进语义交互；提出方面引导的基于对比的多模态语义融合模块，通过对称的mix_up层(Zhang et al., 2018)和多模态融合层来促进跨模态特征融合。通过这种层次化的对齐和融合策略，我们能够更有效地整合图文信息，显著提高面向方面的情感极性识别的准确性。我们的主要贡献如下：

- 本文提出基于方面引导的图文渐进融合的多模态情感分析方法，该方法通过渐进式地对齐和融合图文表示，获得了基于方面的多模态表示，从而提高模型对特定方面情感辨别能力。
- 本文提出了基于对比的跨模态语义交互模块，利用方面作为语义支点，指导图像与文本进行跨模态语义交互，以缩小模态间隙。为了促进跨模态特征融合，我们提出了基于对比的多模态语义融合模块，通过利用对称的双层跨模态交互和对比学习，我们实现了多模态语义的融合。
- 在三个基准数据集上对MABSA任务的实验结果证明了我们提出的模型的有效性和鲁棒性，取得了有希望的结果。

2 相关工作

多模态方面级的情感分析（*Multi-modal Aspect-Based Sentiment Analysis*, MABSA）通过整合图像和文本信息，旨在确定句子中特定方面的情感极性。MABSA领域在自然语言处理（*Natural Language Processing*, NLP）中受到日益增长的关注，研究了多种方法以应对图文方面级情感分析的挑战。早期方法着重于运用注意力机制以捕捉多模态数据之间的相关性。Xu et al. (2019)提出一个交互记忆力网络，旨在捕获多模态数据中的复杂关联。Yu and Jiang (2019)提出了TomBERT预训练语言模型，通过引入特定的模态注意力和跨模态交互模块以增强情感分析的性能。Zhou et al. (2021)采用基于多模态交互层的对抗训练策略，将文本和视觉表示对齐至共享空间以用于MABSA任务。然而，图像与文本之间的固有模态差异使多模态融合经常面临模态匹配噪声的挑战。为应对视文融合中的抗噪声问题，Yu et al. (2020)引入了基于双线性交互的门控模块，以缩减模态间的差距，针对实体敏感的多模态情感分类问题。Zhang et al. (2021)提出两个基于记忆的模块，通过捕获内模态与跨模态特征以解决模态对齐问题。进一步，Yu et al. (2022a)引入一个辅助重构模块，弥合文本与视觉表示间的语义鸿沟，实现了情感分析的最新进展。Khan and Fu (2021)采用基于Transformer的图像标题生成模型，通过增强文本表示并减少模态间的语义鸿沟来提升情感识别性能。Zhao et al. (2022)提出了一个知识增强框架（KEF），利用图像中的形容词-名词对来增强MABSA任务中情感分析性能。Yang et al. (2022)提出了一个面向面部敏感表情到情感文本（FITE）方法，使用面部表情作为情绪线索，增强跨模态分析中的情识别能力。Jia et al. (2023)提出一种情感区域识别与融合网络，利用相关情感区域来促进于视觉与文本的多模态融合。Wang et al. (2024)提出了一种全局-局部特征融合的共同注意力，以分层的方式整合全局-局部多模态特征以提高情感分析性能。

3 方法

本文提出了一种基于方面引导的图文渐进融合的多模态情感分析方法，模型总体架构如图2所示。它主要包含四个网络：图文编码器、基于对比的跨模态语义交互模块、基于对比的多模态语义融合模块及输出模块。

3.1 任务描述

给定一组多模态样本集M，每个样本$c \in M$ 由一句文本s、一张与文本相关的图像v、一个方面目标t 以及一个情感标签y 组成，即$c = \{\mathrm{s}, \mathrm{v}, \mathrm{t}, \mathrm{y}\}$。其中文本$s = (w_1, w_2, w_3, \ldots, w_m)$，由$m$ 个单词组成，而方面目标$t = \{t_1, t_2, \ldots, t_n\}$，由$n$ 个单词组成。情感标签y 分为三类：负面、中性和正面，即$y \in \{\mathrm{negative}, \mathrm{neutral}, \mathrm{positive}\}$。目标是学习一个基于方面的情感分类器，以便它能正确预测未见测试集中的方面的情感极性。

3.2 图文编码器

3.2.1 文本编码器

文本和方面目标的编码器采用预训练的BERT(Devlin et al., 2019)模型。我们分别在文本s和方面术语t 的开始和结束处添加特殊标记[CLS] 和[SEP]，然后将它们输入到预训练的BERT编码器中，以获取文本表示H_s 和方面表示H_t，可以表示为，

$$H_s = \mathrm{BERT}(s) \tag{1}$$

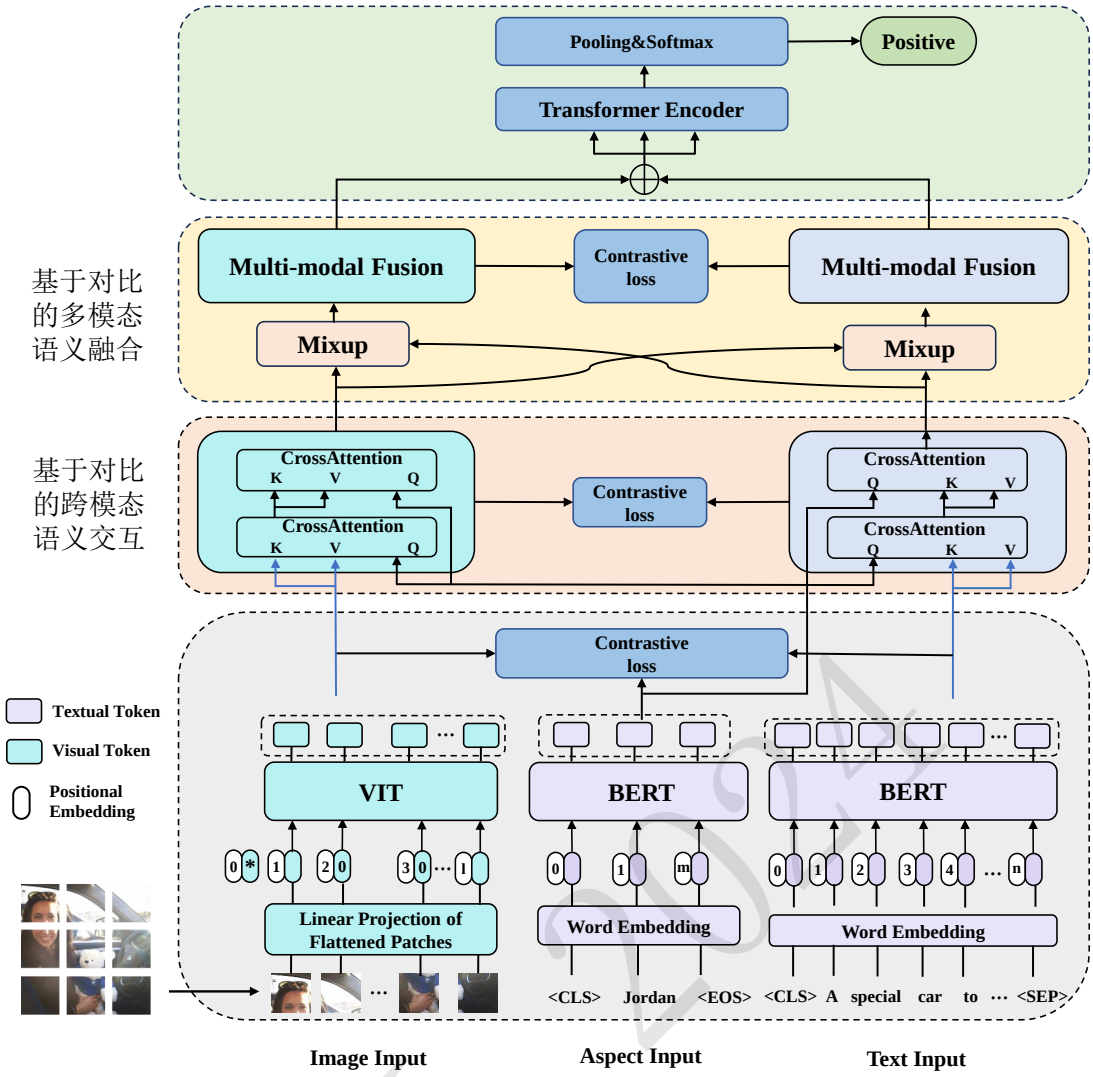

图 2. 基于方面引导的图文渐进融合的多模态情感分析模型

$$H_t = \text{BERT}(t) \tag{2}$$

其中，$H_s \in \mathbb{R}^{m \times d}$, $H_t \in \mathbb{R}^{n \times d}$，$d$ 是隐藏维数，m 和n 分别是输入长度。

3.2.2 图像编码器

图像由预训练的Vision Transformer模型（ViT）(Dosovitskiy et al., 2021)来提取初始特征，首先将图像v 切分成l 个图像块序列$[v_1, v_1, \ldots, v_l] \in \mathbb{R}^{l \times (P^2 \times C)}$，然后在前面加上一个特殊的标记$v_{[\text{CLS}]}$，通过线性投影层$V$，再加上位置嵌入得到图像嵌入$\bar{v}$，可以表示为，

$$\bar{v} = \left[v_{[\text{CLS}]}V, v_1V, v_1V, \ldots, v_lV\right] + V^{\text{pos}} \tag{3}$$

其中，$\bar{v} \in \mathbb{R}^{(l+1) \times d}$，线性投影层$V \in \mathbb{R}^{(P^2 \times C) \times d}$，$V^{\text{pos}} \in \mathbb{R}^{(l+1) \times d}$ 是位置嵌入，d 是维数，(P, P) 是每个图像块的分辨率，C 是图像的通道数。进一步地，我们将$\bar{v}$ 输入到Transformer 中，以获取图像表示$H_v \in \mathbb{R}^{(l+1) \times d}$:

$$H_v = \text{Transformer}(\bar{v}) \tag{4}$$

3.2.3 方面引导的图文对比学习

为了拟合不同模态的初始表示差距，本文采用方面引导的图文对比学习，以方面表征为枢纽，通过在公共语义空间中拉近图像-方面与文本-方面的距离来促进图文语义对齐，具体可表

示如下，

$$L_a = L(H_t, H_s) + L(H_t, H_v) \tag{5}$$

其中，L_a 表示方面引导的图文对比损失，$L(H_t, H_s)$ 表示方面表征和文本表征之间的对比损失，$L(H_t, H_v)$ 表示方面表征和图像表征之间的对比损失，定义如下，

$$L(\alpha, \beta) = \frac{1}{2}(L_{\alpha 2\beta} + L_{\beta 2\alpha}) \tag{6}$$

其中$\alpha, \beta \in \{H_t, H_v, H_s\}$. $L_{\alpha 2\beta}$ 和$L_{\beta 2\alpha}$具体表示如下：

$$L_{\alpha 2\beta} = -\log \frac{\exp(\alpha_i^\top \beta_i/\tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j/\tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j/\tau)} \tag{7}$$

$$L_{\beta 2\alpha} = -\log \frac{\exp(\beta_i^\top \alpha_i/\tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j/\tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j/\tau)} \tag{8}$$

其中，α_i 和β_j 分别是α 中第i 个特征和β 中第j 个特征的归一化表示。N 是批次大小，τ 是温度参数。

3.3 基于对比的跨模态语义交互模块

本模块旨在利用方面表示作为跨模态语义支点，引导文本与图像之间的整体语义交互对齐，进而促进跨模态融合。为达成此目的，本文采用了双层跨模态注意力机制来处理方面-文本与方面-图像间的语义交互。首先，以方面表示H_t 作为查询（Query），图像表示H_v 和文本表示H_s 分别作为键（Key）/值（Value），通过添加残差和层归一化操作(Ba et al., 2016) 获得基于方面的浅层图-文表示H'_v 和H'_s，如下：

$$H'_v = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H_v, H_v)) \tag{9}$$

$$H'_s = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H_s, H_s)) \tag{10}$$

其中，H'_v, $H'_s \in \mathbb{R}^{n\times d}$，LN表示层归一化操作。

进一步地，继续将H_t 视为查询，将H'_v、H'_s 视为键和值，再次通过跨模态注意力机制进行交互，并加入残差及层归一化操作，以获得更深层的基于方面的图-文表示H''_v、H''_s，如下所示：

$$H''_v = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H'_v, H'_v)) \tag{11}$$

$$H''_s = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H'_s, H'_s)) \tag{12}$$

其中，H''_v, $H''_s \in \mathbb{R}^{n\times d}$。接着，对基于方面的图-文表示$H''_v$ 和H''_s 进行对比学习，以进一步促进文本与图像之间的整体语义对齐。对比学习的损失函数表示如下，

$$L_b = L(H''_v, H''_s) = \frac{1}{2}(L_{v2s} + L_{s2v}) \tag{13}$$

其中，L_{v2s} 和L_{s2v} 分别表示从图像到文本和从文本到图像的对比损失，定义如下：

$$L_{v2s} = -\log \frac{\exp(\alpha_i^\top \beta_i/\tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j/\tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j/\tau)} \tag{14}$$

$$L_{s2v} = -\log \frac{\exp(\beta_i^\top \alpha_i/\tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j/\tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j/\tau)} \tag{15}$$

其中，α 和β 分别代表H''_v 和H''_s。

3.4 基于对比的多模态语义融合模块

为了进一步实现多模态语义的融合，我们引入了一个基于对比的多模态融合模块。首先，通过采用mix_up技术进行基于方面的图文模态的平滑过渡，从而得到浅层多模态表示，以文本融合图像的表示H_{s2v} 和以图像融合文本的表示H_{v2s}，具体定义如下：

$$H_{s2v} = w_{\text{mix}} H_s'' + (1 - w_{\text{mix}}) H_v'' \tag{16}$$

$$H_{v2s} = w_{\text{mix}} H_v'' + (1 - w_{\text{mix}}) H_s'' \tag{17}$$

其中，w_{mix} 为融合权重，H_{s2v} 和$H_{v2s} \in \mathbb{R}^{n \times d}$ 代表浅层的多模态表示。

为避免过度融合而丧失图文特有的信息，我们采用并行的带有归一化的多头自注意力层和多头跨模态注意力层，以保留浅层的多模态表示。具体地，将平滑过渡的H_{s2v} 分别输入自注意力层，同时将H_{s2v} 作为查询，H_{v2s} 作为键和值输入跨模态注意力层，通过集成这两者的输出得到最终基于文本的多模态表示$E_t \in \mathbb{R}^{n \times d}$：

$$E_t = \text{LN}\left[(H_{s2v} + \text{Self\_ATT}(H_{s2v})) + (H_{s2v} + \text{Cross\_ATT}(H_{s2v}, H_{v2s}))\right] \tag{18}$$

其中，Self_ATT 是多头自注意力层，Cross_ATT 是多头跨模态注意力层。

类似于E_t，我们将H_{v2s} 输入到自注意力层，将H_{v2s} 视为查询，同时将H_{s2v} 视为键和值输入跨模态注意力层，以此来获取基于图像的多模态表示$E_v \in \mathbb{R}^{n \times d}$：

$$E_v = \text{LN}\left[(H_{v2s} + \text{Self\_ATT}(H_{v2s})) + (H_{v2s} + \text{Cross\_ATT}(H_{v2s}, H_{s2v}))\right] \tag{19}$$

接着，对基于文本的多模态表示E_t 和基于图像的多模态表示E_v 进行多模态对比学习，作为辅助损失函数以进一步促进多模态信息的对齐与融合。损失函数具体表示如下：

$$L_c = L(E_t, E_v) = \frac{1}{2}(L_{v2t} + L_{t2v}) \tag{20}$$

其中，L_{v2t} 和L_{t2v} 分别表示图像到文本和文本到图像的对比损失，定义如下：

$$L_{v2t} = -\log \frac{\exp(\alpha_i^\top \beta_i / \tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j / \tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j / \tau)} \tag{21}$$

$$L_{t2v} = -\log \frac{\exp(\beta_i^\top \alpha_i / \tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j / \tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j / \tau)} \tag{22}$$

其中，α 和β 分别代表E_t 和E_v。

3.5 输出模块

将基于文本的多模态表示E_t 和基于图像的多模态表示E_v 通过最后一维进行拼接，得到多模态表示$E_M = E_t \,\|\, E_v$，并将其输入到一个Transformer编码器中，我们得到：

$$H = \text{Transformer}(E_M, E_M, E_M) \tag{23}$$

其中，$H \in \mathbb{R}^{n \times 2d}$ 是Transformer编码器输出的最终多模态表示，$\|$ 表示拼接操作。

然后，将池化后的H 送入softmax层进行情感分类：

$$P(y \mid H) = \text{softmax}(\text{Pooling}(H)W) \tag{24}$$

其中，$W \in \mathbb{R}^{d \times 3}$，Pooling 表示池化操作。

使用交叉熵损失作为分类损失L 来学习W：

$$L = \frac{1}{N} \sum_{j=1}^{N} P(y^j \mid H) \tag{25}$$

最终的目标函数定义为组合目标函数：

$$L_{\text{all}} = L + \lambda_a L_a + \lambda_b L_b + \lambda_c L_c \tag{26}$$

其中，λ_a, λ_b 和λ_c 是超参数。

4 实验

4.1 数据集

我们在三个基准数据集上进行实验，包括Twitter-15 和Twitter-17 [-1]以及MASAD[0] 数据集。Twitter-15 和Twitter-17 数据集由Yu and Jiang (2019)提出，包含了2014-2015 年和2016-2017 年的多模态推文，这些推文覆盖了诸如政治、体育、娱乐等多种主题。每个多模态推文包括一段文本、文本旁的图片、文本中提及的方面术语，以及每个方面的情感标签。情感标签从负面、中性、正面的集合中分配。MASAD 数据集由Zhou et al. (2021) 提出。该数据集涵盖了七个领域内的57 个预定义方面，包括食品、商品、建筑、动物、人类、植物和景观等。方面的情感标签仅包括正面和负面。表1 和表2 提供了这三个基准数据集的详细信息。

	TWITTER-15			TWITTER-17		
Label	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total	3179	1122	1037	3562	1176	1234

表 1. TWITTER-15和TWITTER-17数据集的详细数据

	Train			Test			Total		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Food	2360	433	2793	592	109	701	2952	542	3494
Goods	2671	1674	4345	743	512	1255	3414	2186	5600
Buildings	1450	970	2420	367	245	612	1817	1215	3032
Animal	3023	2208	5231	1126	670	1796	4149	2878	7027
Human	1999	1838	3837	503	464	967	2502	2302	4804
Plant	2819	2607	5426	1269	947	2216	4088	3554	7642
Scenery	3600	1936	5536	907	490	1397	4507	2426	6933
Total	17922	11666	29588	5507	3437	8944	23429	15103	38532

表 2. MASAD数据集的详细数据

4.2 参数设置和评价指标

本文采用预训练的`bert-base-uncased` 模型来提取文本和方面的初始特征，同时使用预训练的`vit_base_patch16_224` 模型作为图像编码器。这两种模型的隐藏维度均设置为768，多头自注意力和交叉注意力机制中的头数设置为8，前馈神经网络的内层维度设置为2048。为了防止模型过拟合，我们冻结了`vit` 的参数。模型的所有可训练参数使用AdamW 优化器进行更新。对于数据集Twitter-15 和Twitter-17，学习率分别设置为1×10^{-5} 和1×10^{-4}，数据集MASAD学习率设置为1×10^{-5}，批次大小均为32。超参数w_{mix}为0.85，λ_a, λ_b 和λ_c均为0.001。实验在单个RTX 3090 GPU 上进行。

本文采用准确率（Acc）和F1分数（Macro-F1）作为评估指标来衡量本文提出的模型在多模态方面级情感分析任务中的性能。

4.3 基线模型

我们采用三种类型的基线模型来评估方面级情感分析的性能，包括仅图像的方法、仅文本的方法和多模态方法，具体如下：

[-1]https://drive.google.com/file/d/1PpvvncnQkgDNeBMKVgG2zFYuRhbL873g/view
[0]https://drive.google.com/file/d/19YJ8vEYCb-uEKUqSGFmysUTvNzxhVKFE/view

仅图像方法： Res-Aspect(He et al., 2016)：采用Resnet和Bert模型分别提取视觉和方面特征。

仅文本方法： MGAN(Fan et al., 2018)：采用多粒度注意力网络，促进文本与方面间的交互进行情感分类。BERT(Devlin et al., 2019)：预训练的Bert模型，通过输入方面目标和文本进行情感分类。BART(Lewis et al., 2020)：一种序列到序列的预训练BART模型，通过输入句子及其方面目标预测情感极性。ATAE-LSTM(Wang et al., 2016)：结合方面信息并通过注意力机制识别句子显著部分的模型。IAN(Ma et al., 2017b)：设计交互注意力机制以映射方面与相应句子间关系的模型。RAM(Chen et al., 2017)：引入加权记忆机制以精确捕捉指定方面相关情绪信息的模型。TNet(Li et al., 2018)：利用CNN层从Bi-RNN派生的转换词表示中提取关键特征的模型。

多模态方法： Res-MGAN：设计跨模态多粒度注意力网络，促进文本与图像交互以进行情感分类。Res-Bert：使用预训练的BERT和Resnet模型提取文本和视觉特征，并应用跨模态注意力机制进行多模态融合以进行情感分类。MIMN(Xu et al., 2019)：提出多跳记忆网络捕捉文本与视觉模态间交互的模型。ESAFN(Yu et al., 2020)：提出实体敏感的多模态注意力融合网络，动态捕获方面-文本与方面-图像表示的模型。ViLBERT(Lu et al., 2019)：基于图像与文本训练的多Transformer模型ViLBERT。TomBERT(Yu and Jiang, 2019)：提出面向目标的多模态BERT架构，捕获方面感知的多模态表示，增强多模态情感分类性能。CapTrBERT(Khan and Fu, 2021)：使用BERT通过图像标题和句子作为输入判断情感极性。MMAP(Zhou et al., 2021)：采用对抗训练与多模态层合并文本和图像进行情感分析。saliencyBERT(Wang et al., 2021)：使用递归注意力网络捕获方面敏感的视觉特征。HIMT(Yu et al., 2022a)：使用辅助重构模块的层次交互模型弥合文本和视觉表示间的语义差距。KEF(Zhao et al., 2022)：采用知识增强框架（KEF）从图像中提取形容词-名词对，改善视觉注意力与情感预测。FITE(Yang et al., 2022)：一种应用面部指示文本情绪（FITE）方法，利用面部表情增强跨模态情感识别。ARFN(Jia et al., 2023)：一种情感区域识别与融合网络（ARFN）方法，通过识别和利用图像中的情感区域，增强多模态方面级情感分类。GLFFCA(Wang et al., 2024)：一种基于全局-局部特征融合的共同注意力（GLFFCA）的多模态方面级情感分析方法。

4.4 在TWITTER-15和TWITTER-17数据集上的实验结果

我们在两个MABSA基准数据集上进行比较实验，以展示我们提出方法的有效性，如表3所示。我们可以看到：1) 两个基准数据集上的实验结果显示，我们提出的多模态方法优于大部分现有方法，与次优模型相比F1值在两个数据集上分别提高了0.1%和0.57%，这说明了我们提出的方法的有效性。2)与单模态方法相比，本文提出的方法显著优于所有单模态方法上，例如仅文本的方面级情感分析方法和仅图像的方面级情感分析方法。实验结果证明了仅文本的方法可以有效的利用视觉模态来提高情感分析的性能。3)与其他多模态方面级情感分析方法相比，我们提出的方法在所有评估指标上都比最近的多模态融合方法有了显着改进，这表明方面引导的语义对齐和多模态融合有助于提高情感分析性能，这确认了我们提出方法的优越性。

4.5 在MASAD数据集上的实验结果

为了进一步证明本文方法的鲁棒性，我们在新发布的MASAD数据集上进行了实验。表4中的实验结果显示：1) 本文的方法取得了最好的效果，在准确率ACC和F1分数上均显著超过其他多模态方面级情感分析方法。2) 在MASAD数据集的七种方面术语分类上，本文方法的结果全部优于其他多模态方面级情感分析方法。在MASAD数据集上的实验结果进一步证明了我们提出方法的有效性和鲁棒性。此外，值得注意的是，MASAD数据集是一个最近才发布的数据集，因此只有少数MABSA方法在此数据集上进行了情感分析性能评估。

4.6 消融实验

4.6.1 模型不同组件的消融分析

为了进一步验证本文提出方法的有效性，我们在TWITTER-15和TWITTER-17两个数据集上进行了不同组件的消融实验，消融结果如表5所示。实验结果表明：1)移除方面引导的图文对比学习的模型，相比于完整的模型，在两个数据集上的两个评估指标都略微下降，这表明方面引导的图文对比学习一定程度上弥合了图文模态的差异。2)移除基于对比的跨模态语义交互模块的模型，相比于完整的模型，在两个数据集上的两个评估指标都有所下降，这表明方面引

Modality	Methods	TWITTER-15		TWITTER-17	
		Acc	Macro-F1	Acc	Macro-F1
Image	Res-Aspect	59.49	47.79	57.86	53.98
Text	MGAN	71.2	64.2	64.8	61.5
	BERT	74.3	70.0	68.9	66.1
	BERT+BL	74.3	70.0	68.9	66.1
	BART	76.0	67.6	69.5.	67.0
Text+Image	Res-MGAN	71.7	63.9	66.4	63.0
	Res-Bert	75.02	69.21	69.2	66.48
	MIMN	71.84	65.69	65.88	62.99
	ESAFN	73.38	67.37	67.83	64.22
	ViLBERT	73.76	69.85	67.42	64.87
	TomBERT(resnet)	76.60	71.57	69.42	67.70
	TomBERT(faster-rcnn)	77.03	72.85	69.77	67.59
	CapTrBERT	78.01	73.25	69.77	68.42
	saliencyBERT	77.03	72.36	69.69	67.59
	HIMT	78.41	73.68	71.14	69.16
	KEF-saliencyBERT	78.15	73.54	71.88	68.94
	FITE	78.49	73.90	70.90	68.70
	TMGM	74.13	67.93	67.74	65.32
	AFRN	78.50	73.70	70.58	68.43
	GLFFCA	77.72	74.21	71.15	69.45
	Our model	**77.92**	**74.31**	**71.22**	**70.02**

表 3. 在TWITTER-15和TWITTER-17数据集的比较结果

Modality	Methods	Food		Goods		Buildings		Animal		Human		Plant		Scenery		Average	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Text	ATAE-LSTM	93.06	86.13	95.31	94.85	94.32	94.64	93.56	92.85	91.55	91.34	93.05	93.4	91.87	91.03	93.25	92.03
	IAN	93.56	88.3	95.77	95.3	94.78	94.98	94.05	93.32	92.04	92.13	93.85	93.71	92.33	91.87	93.77	92.8
	RAM	93.64	89.25	95.59	95.12	94.57	94.71	94.29	93.55	91.75	91.87	93.5	93.37	92.15	91.39	93.64	92.75
	TNet	94.34	90.18	95.88	95.34	95.04	94.9	94.76	93.87	92.27	92.16	94.38	94.21	92.76	91.98	94.2	93.23
Text+Image	MIMN	94.72	91.39	95.93	95.87	96.26	95.8	95.03	94.06	92.54	92.31	95.04	94.97	93.17	92.38	94.67	93.83
	TomBERT	95.56	91.8	96.05	96.1	96.53	96.04	95.62	94.78	92.67	94.97	95.67	95.3	93.63	92.94	95.1	94.21
	MMAP	95.75	92.89	96.55	96.44	96.86	96.85	95.92	95.62	92.74	92.74	97.02	96.97	94.57	94.15	95.63	95.09
	Our model	**96.43**	**95.81**	**97.57**	**97.06**	**97.32**	**97.29**	**99.36**	**99.56**	**96.84**	**96.75**	**97.48**	**97.59**	**95.34**	**94.37**	**97.19**	**96.92**

表 4. 在MASAD数据集的比较结果

导的的跨模态语义交互可以为情感分析提供更精确的信息。3)移除移除基于对比的多模态语义融合模块，相比于完整的模型，在两个数据集上ACC和F1评估指标都大幅下降，这表明方面引导的跨模态语义交互有助于指导图文多模态对齐融合，提升情感分析的性能。通过以上的消融实验对比分析，验证了本文模型不同组件的有效性。

4.6.2 视觉信息的消融分析

为了验证本文提出方法中视觉信息的有效性，我们通过将与文本相关的图像替换为空白图像或随机选取的图像来进行实验，实验结果如表6 所示。实验结果表明，在TWITTER-15和TWITTER-17数据集上，当我们将与文本相关的图像替换为空白图像或随机选取的图像时，情感分类性能显著下降。这证明了我们提出的多模态融合方法的有效性。

Method	TWITTER-15		TWITTER-17	
	Acc	Macro-F1	Acc	Macro-F1
Full model	**77.92**	**74.31**	**71.22**	**70.02**
w/o Contrastive	77.46	73.68	69.77	68.58
w/o Interaction	77.53	72.66	68.80	68.13
w/o Fusion	76.39	71.60	67.62	67.64

表 5. 在TWITTER-15和TWITTER-17数据集上模型不同组件的消融实验，w/o Contrastive是移除方面引导的图文对比学习的模型，w/o Interaction是移除基于对比的跨模态语义交互模块的模型，w/o Fusion 是移除基于对比的多模态语义融合模块的模型，Full model是本文完整的模型。

Method	TWITTER-15		TWITTER-17	
	Acc	Macro-F1	Acc	Macro-F1
	77.92	**74.31**	**71.22**	**70.02**
The blank image	77.34	71.78	68.59	67.75
The random image	76.93	72.23	67.47	67.43

表 6. 在TWITTER-15和TWITTER-17数据集上视觉信息的消融实验

4.7 实例分析

Image and text	 (a)**Happy New Year**! Watching San Jose - Anaheim replay ... believe it or not, this was the scene in **Riverside County** today.	 (b)Embattled Metro Councilman **Dan Johnson** to debate challenger **John Wit**, an independent. # Louisville	 (c) **The Seth Leibsohn Show** TONIGHT: Attorney Sheila Polk on Legalizing # Marijuana amp **Steve Hayward** !
Ground truth	(New Year, Pos) (Riverside County, Neu)	(John Wit, Neu) (Dan Johnson, Neu)	(The Seth Leibsohn Show, Neu) (Steve Hayward, Pos)
w/o Contrastive	(Pos ✓, Neu ✓)	(Neu ✓, Neu✓)	(Neu✓, Neg✗)
w/o Interaction	(Pos ✓, Neg ✗)	(Neg✗, Neu✓)	(Neu✓, Neg✓)
w/o Fusion	(Pos , Neg ✗)	(Neu ✓, Neg✗)	(Neu✗, Neu✗)
Full method	(Pos ✓, Neu ✓)	(Neu✓, Neu✓)	(Neu✓, Pos✓)

图 3. MABSA数据集上的几个案例研究。其中，Neu、Neg 和Pos 分别代表中性、负面和正面情感。符号✔ 和✗ 分别表示正确和不正确的情感极性预测。

图3展示了Twitter-15数据集中的三个示例。我们可以看到：1）本文的完整模型准确预测了这三个示例中各个方面的情感极性。2）当移除方面引导的图文对比学习时，(c)中的Steve Hayward方面情感极性预测出现错误。这一结果证实了方面引导的图文对比学习起到了模态对齐的作用，提高了MABSA的情感分析性能。3）当移除基于对比的跨模态语义交互模块时，(a)中的Riverside County方面和(b)中的John Wit方面的情感极性预测出现错误。这一结果证实了基于对比的跨模态语义交互模块很好地进行了语义的交互对齐，提高了MABSA的情感分析性能。4）移除基于对比的多模态语义融合模块导致(a)中的Riverside County方面、(b)中

的Dan Johnson方面以及(c)中的The Seth Leibsohn Show方面和Steve Hayward方面的情感极性预测出现错误，这证明了基于对比的多模态语义融合模块很好的融合了面向方面的多模态特征，提高了MABSA的情感分析性能。案例研究进一步验证了我们提出的方法在多模态方面级情感分析任务中的有效性。

5 结论

本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，通过方面引导的跨模态语义交互和多模态语义融合来缓解图像和文本之间的模态差异，实现图文多模态的对齐和融合，提高多模态特征的判别性。与现有方法相比，在三个基准任务上取得了强有竞争力的结果，三个基准数据集上的实验结果证明了本文提出的方法的有效性和优越性。深入的分析了所提出的模块的贡献，并证明了所提出方法用于多模态方面级情感分析的有效性。在未来的工作中，我们会重点探索如何深度挖掘图像中更多有效的信息，以进一步提升多模态方面级情感分析的性能。

参考文献

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3433–3442.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Li Jia, Tinghua Ma, Huan Rong, and Najla Al-Nabhan. 2023. Affective region recognition and fusion network for target-level multimodal sentiment classification. *IEEE Transactions on Emerging Topics in Computing*, pages 1–11.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, July. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017a. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4068–4074. AAAI Press.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017b. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4068–4074.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, page 70–77, New York, NY, USA. Association for Computing Machinery.

Huy Thanh Nguyen and Minh Le Nguyen. 2018. Effective attention networks for aspect-level sentiment classification. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 25–30.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 – November 1, 2021, Proceedings, Part III*, page 3–15, Berlin, Heidelberg. Springer-Verlag.

Shunjie Wang, Guoyong Cai, and Guangrui Lv. 2024. Aspect-level multimodal sentiment analysis based on co-attention fusion. *International Journal of Data Science and Analytics*, pages 1–14, 01.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 371–378.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Kai Chen, and Rui Xia. 2022a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–1.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022b. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4482–4488.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, 24(6):1957–1974.

Fei Zhao, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6784–6794, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58.