计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):253−260 [doi: 10.15888/j.cnki.csa.009871] [CSTR: 32024.14.csa.009871]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于对齐优化的多模态讽刺检测①

曾碧卿[1,2], 陈威海[1]

[1](华南师范大学 人工智能学院, 佛山 528225)
[2](华南师范大学 阿伯丁数据科学与人工智能学院, 佛山 528225)
通信作者: 曾碧卿, E-mail: zengbiqing137@163.com

摘　要: 讽刺是一种修辞手法, 通过言辞或行为表达出与字面意义相反或不同的含义, 常用于批评、讽刺、幽默或反讽, 通常包含对某种情况或观点的嘲笑或挖苦. 由于讽刺的复杂性, 导致讽刺检测很难只通过文本单个模态进行. 因此, 多模态讽刺检测得到了更多研究者的关注. 现有的方法通过注意力机制进行多模态讽刺检测, 然而它们在对齐和融合阶段有所不足, 无法筛选出对齐信息中的重要信息从而影响模型性能. 本文提出了一个基于注意力和图注意力的模型来进行多模态讽刺检测, 它通过多头跨模态注意力模块进行对齐, 通过自注意力增强两个模块输出中的重要信息的表达. 该模型的效果在一个基于 Twitter 的公开讽刺检测数据集上得到了验证.

关键词: 多模态讽刺检测; 自注意力机制; 对齐; 跨模态注意力机制; 图注意力机制

引用格式: 曾碧卿,陈威海.基于对齐优化的多模态讽刺检测.计算机系统应用,2025,34(7):253–260. http://www.c-s-a.org.cn/1003-3254/9871.html

Multimodal Sarcasm Detection Based on Alignment optimization

ZENG Bi-Qing[1,2], CHEN Wei-Hai[1]

[1](School of Artificial Intelligence, South China Normal University, Foshan 528225, China)
[2](Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan 528225, China)

Abstract: Sarcasm is a rhetorical technique that expresses meanings opposite to or different from literal meanings through words or behaviors, often used for criticism, satire, humor, or irony, and typically involves mockery or ridicule of certain situations or viewpoints. Due to the complexity of sarcasm, it is difficult to detect it through text alone. Therefore, multi-modal sarcasm detection has received increasing attention from researchers. Existing methods use attention mechanisms for multi-modal sarcasm detection. However, they exhibit limitations in alignment and fusion stages, which ultimately compromises model performance. This study proposes a model based on both attention and graph attention for multi-modal sarcasm detection. This model employs a multi-head cross-modal attention module for alignment and utilizes self-attention to enhance the representation of critical information in the output of both two modules. The effectiveness of this model has been validated on a public dataset for sarcasm detection based on Twitter.

Key words: multi-modal sarcasm detection; self-attention mechanism; alignment; cross-modal attention mechanism; graph attention mechanism

1　引言

讽刺是一种通过批评、讽刺或幽默的方式来表达对某个人、群体、观点或情况的不满或不赞同的文学或言论手法. 通常, 讽刺的目的是揭示或批判某种荒谬、虚伪、愚蠢或不合理的现象, 以引起读者或听众的思考或反思. 最近, 在社交媒体中, 讽刺这种现象几

① 基金项目: 国家自然科学基金 (62076103); 广东省基础与应用基础研究基金 (2021A1515011171); 广州市基础研究计划 (202102080282)
收稿时间: 2024-11-25; 修改时间: 2024-12-17; 采用时间: 2025-01-07; csa 在线出版时间: 2025-05-27
CNKI 网络首发时间: 2025-05-28

乎随处可见. 然而, 由于讽刺的多样性和微妙性, 讽刺检测迄今为止都是一项具有挑战性的任务.

多模态讽刺检测的一大挑战是正确理解多个模态中包含的信息, 从而推导出作者试图传达的实际含义. 例如, 仅解读图 1 的标题可能会认为作者想要表达对星巴克提供一杯咖啡的感激之情 (一种积极的情感表达). 然而, 一旦将图片内容与文本关联起来, 就可以轻易得出作者对咖啡仅为 2/3 杯的容量感到不满的含义. 因此, 为了正确检测讽刺, 有必要考虑所有方面 (即多模态), 以有效捕捉所有情感表达, 并识别出它们之间的不一致.

图 1 多模态讽刺检测的一个例子, 英文文本为: “Starbucks thanks for my 2/3 tall cafe latte”

之前的研究中已经提出了几种方法用于检测多模态数据中的讽刺信息 (例如, 包含文本和图像的数据). 例如, Wang 等[1]提出了一种利用注意力机制来识别模态内不一致性和模态间不一致性的模型. Liang 等[2,3]探索了使用图神经网络来识别文本和图像模态之间复杂的不一致性. Pramanick 等[4]在多模态讽刺检测中采用自注意力机制来处理模态内和模态间的对应关系. Liu 等[5]通过捕捉原子级别和组合级别的不一致性来探索讽刺检测. 然而, 据我们所知, 之前的研究只是对不同模态的信息进行对齐, 没有进一步优化对齐, 从而筛选出更为重要的对齐信息.

为了研究优化对齐多模态信息是否有助于讽刺检测, 本文提出一种四模块神经网络架构, 如图 2 所示. 图 2 架构中两个模块专门用于信息对齐. 第 1 个模块使用两个编码器分别从原始文本和图像数据中提取特征. 第 2 个模块使用跨模态注意力网络和自注意力网络将图像信息融合到文本特征中并进行对齐, 从而生成第 1 组增强的文本特征. 增强的文本特征随后与图像特征进行乘法运算 (点积), 以获得对齐信息. 第 3 个模块进一步使用图神经网络和图神经注意力网络对文本和图像信息进行对齐. 它将从第 2 个模块获得的增强文本特征和图像特征作为输入, 分别使用图网络和图注意力网络生成每个特征的新表示 (特征向量). 这两个新获得的特征向量再进行乘法运算 (点积), 以获取文本和图像信息的更深层次对齐信息. 第 4 个模块使用来自第 2、3 模块的对齐信息 (点积) 来预测是否存在讽刺. 它首先将第 2、3 模块的点积结果进行串联, 然后通过全连接层和 *Softmax* 层生成最终预测.

本文的主要贡献有以下 3 点.

(1) 探索了使用自注意力机制优化对齐后的文本和图像信息, 从而增加了在多个模态中发现讽刺对比的可能性.

(2) 使用了一种双重对齐神经网络架构, 用于融合文本和图像信息. 其中一种方式依赖于跨模块注意力, 另一种方式依赖于图网络和图注意力网络.

(3) 在一个公共数据集上比较了提出的讽刺检测方法与其他方法, 结果表明本文方法的性能在讽刺检测任务上有所提升.

2 相关工作

2.1 单模态讽刺检测

早期的讽刺检测研究主要集中在文本中的对比发现上[6,7], 因为大多数讽刺都是以文本形式表达的. Kim[6]使用卷积神经网络 (CNN) 进行讽刺检测, 并通过微调词向量改进性能; Tay 等[7]提出了一种基于注意力机制的神经网络模型, 该模型注重内在对比而不是跨界对比, 使其能够明确地建模对比和不协调之处. 由于讽刺常与其他非语言线索一起使用, 以提醒人们不要只对字面意思进行理解, 如语调、面部表情和肢体语言等. 忽视或误解这些线索可能会导致对说话者意图的错误解读.

2.2 多模态讽刺检测

近年来, 随着社交媒体每天产生大量的多模态数据 (例如文本、图像、视频), 研究重点逐渐转向多模态检测. 在一项研究中, Schifanella 等[8]定义了多模态讽刺检测任务, 并提出通过连接文本和图像信息的表示 (特征) 向量来结合这两种信息. Cai 等[9]使用 Twitter

数据构建了一个多模态讽刺检测数据集, 该数据集已成为这一领域中最受关注的数据集之一. 除了数据工作, Cai 等还提出了一种基于双向 LSTM 的层次融合模型用于多模态讽刺检测. Xu 等[10]探索了使用分解和关系网络进行比较和语义关联. Wang 等[1]开发了一种利用注意力机制识别模态内和模态间不一致性的模型. Liang 等[2]研究了使用图神经网络来建模文本和图像之间更复杂的不一致性. Pramanick 等[4]采用自注意力机制处理模态内和模态间的对应关系. Liu 等[5]研究了在原子级别和组合级别这两种级别上检测讽刺.

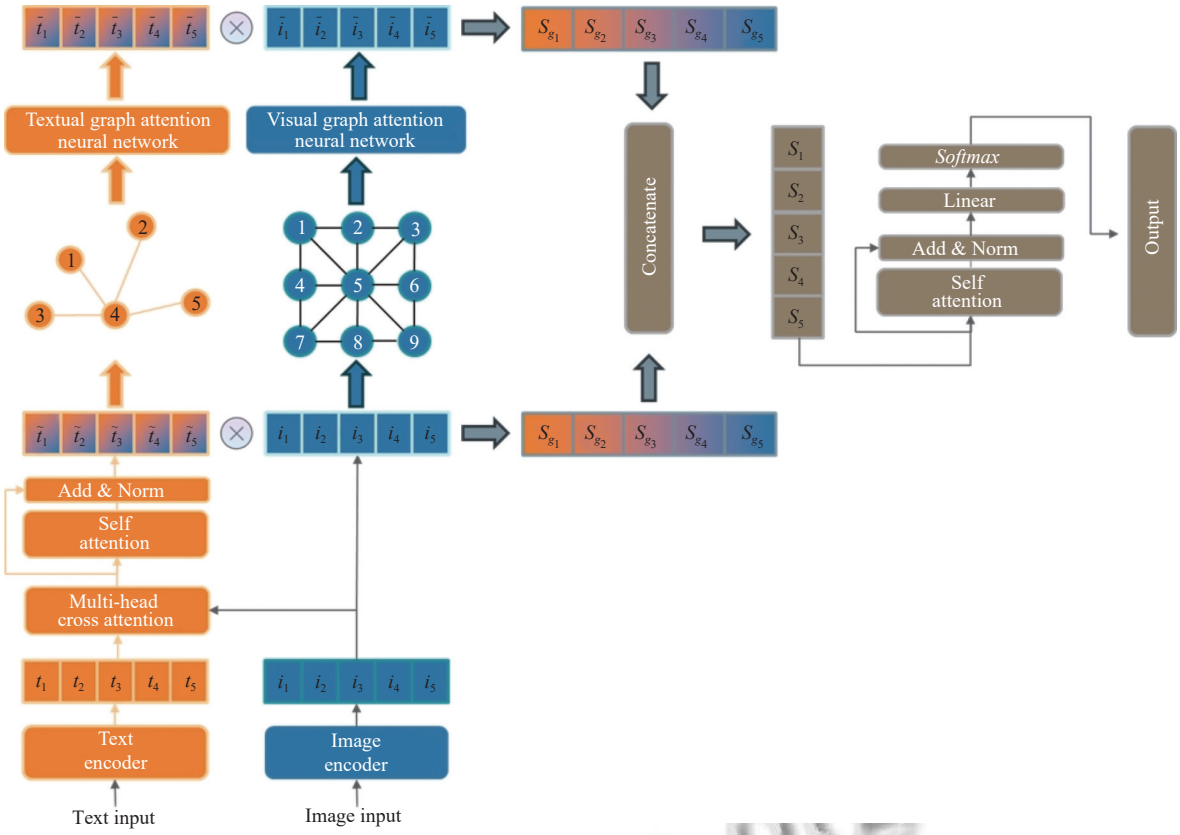

图 2 模型概览图

3 模型及方法

本文模型由 4 个部分组成 (如图 2 所示).

(1) 特征提取模块: 该模块利用编码器从原始数据中提取文本和图像特征.

(2) 注意力对齐模块: 该模块采用跨模块注意力层来结合文本和图像信息.

(3) 图对齐模块: 该模块应用图注意力神经网络以进一步合并文本和图像信息.

(4) 预测模块: 该模块使用多个层来预测讽刺的存在.

3.1 特征提取模块

特征提取模块以图像-文本对作为输入. 然后, 原始文本数据被编码为数值向量 $T=[t_1,t_2,\cdots,t_n]$, 其中 $T\in\mathbb{R}^{n\times d}$, 使用 BERT 模型[11]. 原始图像数据也被编码为数值向量. 对于图像模态, 遵循文献[9], 对于每个图像, 我们将其调整大小为 224×224 像素, 将其分成一系列 r 个补丁 $\{p_1,p_2,\cdots,p_r\}$, 并使用 ViT[12]模型将其编码为向量 $I=[i_1,i_2,\cdots,i_r]$, 其中 $I\in\mathbb{R}^{r\times d}$.

3.2 注意力对齐模块

由于两种模态之间的差距, 该模型使用多头跨模态注意力机制对两个不同模态进行对齐, 并使用一个自注意力模块强化向量中的不一致性表达, 第 n 层可以表示为:

$$T_n=norm(T_{n-1}+MCA(T_{n-1},I)) \quad (1)$$

其中, $norm$ 表示层归一化操作, MCA 表示多头跨模态

注意力模块, T_n 是第 n 层的文本特征向量, T_0 是 BERT 得到的初始文本特征向量, 是 ViT 得到的初始图像特征向量. 单层多头跨模态注意力模块的第 i 个注意力头可以表示为:

$$head_i = Softmax\left(\frac{\left(TW_q^i\right)^{\mathrm{T}}}{\sqrt{d/h}}(IW_k^i))W_v^i\right) \tag{2}$$

其中, $I \in \mathbb{R}^{r\times d_I}$ 和 $T \in \mathbb{R}^{n\times d_T}$ 分别是文本特征向量和图像特征向量. $W_q^i \in \mathbb{R}^{d_I\times \frac{d_I}{h}}$、$W_k^i \in \mathbb{R}^{d_I\times \frac{d_I}{h}}$ 和 $W_v^i \in \mathbb{R}^{d_I\times \frac{d_I}{h}}$ 分别是 $head_i \in \mathbb{R}^{n\times \frac{d}{h}}$ 的查询、键和值投影矩阵. 将所有头的输出串连到一起通过一个线性层之后经过一个残差连接层, 最后再经过一个归一化层得到更新后的文本表示:

$$\tilde{T} = norm(T + concat[head_1, head_2, \cdots, head_n]) \tag{3}$$

其中, $norm$ 表示层归一化操作, $concat$ 表示拼接操作. 然后将得到的文本特征向量通过一个自注意力层以强化文本特征向量中不一致性的表达, 之后经过一个残差连接层得到最后的文本特征向量:

$$\tilde{T} = norm(\tilde{T} + SA(T)) \tag{4}$$

其中, SA 代表自注意力模块. 为了与后面图注意力部分的输出进行融合, 该模型采用内积 $Q_a = \frac{1}{\sqrt{d}}\left(\tilde{T}I^{\mathrm{T}}\right)$, 其中, $Q_a \in \mathbb{R}^{n\times r}$ 是文本特征向量和图像特征向量的注意力矩阵. 之后, 文本特征向量将被输入到一个线性层, 再通过 $Softmax$ 激活函数得到最终的注意力分数 S_a:

$$S_a = Softmax(\tilde{T}W_a + b_a)Q_a \tag{5}$$

其中, $W_a \in \mathbb{R}^{d\times 1}$ 和 $b_a \in \mathbb{R}^n$ 是线性层的可训练参数. $S_a \in \mathbb{R}^r$ 为跨模态注意力对齐层输出的注意力分数.

3.3 图对齐模块

该模型通过图注意力层让模型学习到文本模态和图像模态更加复杂的信息. 在具体实现中, 模型首先对输入的文本信息和图像信息分别构建相应的文本图和视觉图. 对于文本模态, 该模型使用 spaCy 工具包为有依赖关系的 token 构建边. 具体来说, 文本图的点是文本中的 token, 而文本图中的边是 spaCy 工具包提取的文本依赖关系, 这种做法已被证明对于各种图相关任务[3,5]是有效的. 对于图像模态, 模型直接通过图像 patch 之间的几何关系来构建视觉图.

然后, 模型通过图注意力网络 (GAT) 得到文本图和视觉图, 用以建模文本之间的语法依赖关系和图像 patch 之间的关系, 后续模型将捕捉这两种关系的不一致性. 在这里, 本文将文本图作为例子进行表示.

$$\alpha_{i,j}^l = \frac{\beta_{i,j}}{\sum_{x=1}^{x=m}\sum_{y=1}^{y=n}\beta_{x,y}} \tag{6}$$

$$\beta_{i,j} = \exp(LeakyReLU(V_h^{\mathrm{T}}[\theta_l t_i^h; \theta_l t_j^h])) \tag{7}$$

$$t_i^{h+1} = \alpha_{i,i}^h \theta_h t_i^h + \sum_{j=1}^{j=n}\alpha_{i,j}\theta_h t_j^h \tag{8}$$

其中, $\theta_h \in \mathbb{R}^{d\times d}$ 和 $V_h \in \mathbb{R}^{2d}$ 是第 h 层 GAT 层的可学习参数. $\alpha_{i,j}^h$ 是注意力得分. t_i^h 表示第 h 层第 i 个节点, $t_i^0 = \tilde{t}_i$ 由对齐后的文本特征向量初始化.

在某些情况下, 由于句子中的单词较少, 模型可能无法构建一个可靠的文本图. 因此, 本文参考文献[5], 通过加入句子特征向量 c 来改善这个问题, 其中 c 可以表示为:

$$c = Softmax(TW_c + b_c)^{\mathrm{T}}\tilde{T} \tag{9}$$

其中, $c \in \mathbb{R}^d$、$W_c \in \mathbb{R}^{d+1}$ 和 $b_c \in \mathbb{R}^n$ 是线性层中可学习的参数. 类似地, 可以得到视觉图的图特征向量 $\hat{I} \in \mathbb{R}^{r\times d}$.

最后, 模型得到文本图向量和视觉图向量之间的注意力分数 S_g, 计算方式类似于跨模态注意力部分:

$$S_g = Softmax([\hat{T}; c]W_g + b_g)^{\mathrm{T}}Q_g \tag{10}$$

其中, $Q_g = \frac{1}{\sqrt{d}}\left(\left[\hat{T}; c\right]\hat{I}^{\mathrm{T}}\right)$ 为文本模态和视觉模态的图注意力矩阵, $W_g \in \mathbb{R}^{d\times 1}$ 和 $b_g \in \mathbb{R}^{n+1}$ 为线性层的可训练参数, $S_g \in \mathbb{R}^r$ 为图注意力模块输出的注意力分数.

3.4 预测模块

最终, 跨模态注意力得到的注意力分数 S_a 和图注意力得到的注意力分数 S_g 将被拼接起来得到最终的预测结果:

$$P_v = Softmax(IW_v + b_v) \tag{11}$$

$$y' = P_v \odot S_a; P_v \odot S_g \tag{12}$$

其中, $P_v \in \mathbb{R}^r$ 是图像模态的权重向量. $W_v \in \mathbb{R}^{d\times 1}$ 是可训练参数. 之后 y' 通过自注意力模块、全连接层和归一化层得到最终结果:

$$y = Softmax\left(W_y\left[norm(y' + SA(y'))\right] + b_y\right) \tag{13}$$

其中, $b_v \in \mathbb{R}^r$、$W_y \in \mathbb{R}^{2\times 2r}$ 和 $b_y \in \mathbb{R}^2$ 是可训练参数, SA 代表自注意力模块.

4 实验

4.1 数据集

本文在多模态讽刺检测 (MSD) 数据集[1]上评估模型方法, 这是唯一一个用于多模态讽刺检测的公共基准数据集. Cai 等[9]使用 Twitter 数据构建了这个数据集, 其中带有讽刺的图像-文本对被标记为正例, 没有讽刺的被标记为负例. 表 1 对该数据集进行了统计. 带有讽刺、挖苦、转发、反讽、笑话、幽默等内容的条目被丢弃. 同时, 带有 URL 的条目也被移除, 以确保完整的信息包含. 此外, 经常与讽刺推文一起出现的推文也被排除在外, 因为它们可能传达讽刺的想法 (例如, 笑话、幽默、抱怨). 以 8:1:1 的比例随机将数据集分为训练集、验证集和测试集. 对验证集和测试集的标签正确性进行了双重检查.

表 1 MSD 数据集的统计信息

类别	训练集	验证集	测试集
讽刺样例	8642	959	959
非讽刺样例	11174	1451	1450
总和	19816	2410	2409

4.2 基线模型

使用预训练的 BERT-base-uncased 模型作为文本特征提取模型, 并使用 ViT 进行视觉特征提取. 通过线性层将这两个特征的维度调整为 200. 模型的其他超参数可见表 2.

表 2 超参数

参数	值
文本最大长度	100
MCA层数	20
MCA Head数量	5
GAT层数	5
MCA中SA层数	2
预测模块SA层数	1
批大小	32
学习率	2E−5
权重衰减	5E−3
Dropout率	0.5

为了验证实验的有效性, 模型与其他几个基线模型进行了比较, 包括文本模型、图像模型和多模态模型, 下面是每个基线模型的简单介绍.

(1) 文本模型

TextCNN[6]: 基于卷积神经网络的文本分类模型. 它利用卷积层提取文本特征, 并应用池化层进行压缩和特征选择. 在所有层之后, 使用全连接层进行分类. TextCNN 在自然语言处理领域的文本分类任务中表现出良好的性能.

Bi-LSTM[13]: 作为循环神经网络 (RNN) 的一种变体, Bi-LSTM 能够处理时间序列数据, 如文本和语音. 通过前向和后向层, Bi-LSTM 能够考虑每个序列项之前和之后的信息, 使模型能够考虑更多的上下文.

SMSD[14]: 一种序列到序列的模型, 使用软解码以概率方式生成输出序列, 允许一定程度的随机性.

BERT[11]: 使用双向编码器表示来自 Transformer 的模型. 如其名所示, 这是一个基于 Transformer 架构的模型, 它在所有层中考虑了左右上下文.

(2) 图像模型

ResNet[15]: 一种深度残差网络结构, 使用“残差块”来解决梯度消失的问题.

ViT[12]: vision Transformer (ViT) 是基于 Transformer 架构的图像分类模型. 它将图像表示为序列而不是像素网格, 并使用 Transformer 的注意机制处理图像的局部和全局信息.

(3) 多模态模型

HFM[9]: 该模型首先提取图像特征向量和属性特征向量, 然后使用属性特征向量和双向 LSTM 网络提取文本特征. 最后, 3 种模态的特征被重构并融合成一个特征向量进行预测.

D&R Net[10]: 它使用卷积神经网络来表示图像和文本, 并通过跨模态比较模块学习文本和图像之间的差异. 然后, 一个语义关联模块将文本和图像特征映射到一个共享的语义空间.

Att-BERT[16]: 该模型通过 3 个步骤来检测讽刺: 1) 分别学习文本和图像的独立表示; 2) 检测每个模块内部的不一致性; 3) 识别跨模块的不一致性.

InCrossMGs[2]: 一种使用图形进行讽刺检测的方法, 具有 4 个步骤: 1) 分别学习文本和图像的独立表示; 2) 将每个表示转换为使用边表示不同元素之间关系的图; 3) 构建一个模拟不同模态之间语义关系的跨模态图; 4) 使用图卷积神经网络 (GCN) 先传播和聚合特征, 然后预测讽刺.

CMGCN[3]: 一种基于图的方法. 类似于 InCrossMGs, 它首先学习文本和图像的表示, 然后使用跨模态图来建模不同模态之间的关系. 之后, 使用图卷积网络传播和聚合跨模态图的特征. CMGCN (不使用外部知识)

被选择为本研究的基线之一.

HKE[5]: 该方法首先表示文本和图像, 然后使用分层方法在多个级别对文本和图像之间的不一致性进行建模. 之后, 使用知识增强方法增强模型对讽刺的理解. HKE (不使用外部知识) 被选择为本研究的基线之一.

4.3 主要结果

本文模型通过与基线模型进行比较来评估其有效性. 如表 3 所示, 多模态模型的表现优于所有单模态模型, 表明使用多模态数据的优势和必要性. 因此, 我们在模型中并入了文本和图像信息, 并使用跨模态注意力对齐它们. 如预期的那样, 我们的模型在一定程度上优于所有基线模型 (包括单模态和多模态).

表 3 模型结果与其他基线模型的比较 (%)

模态	模型	准确率	精确率	召回率	*F*1
文本模态	TextCNN*	80.03	74.29	76.39	75.32
	Bi-LSTM*	81.90	76.66	78.42	77.53
	SMSD*	81.90	76.66	78.42	77.53
	BERT*	83.85	78.72	82.27	80.22
图像模态	ResNet*	64.76	54.41	70.80	61.53
	ViT*	67.83	57.93	70.07	63.43
文本-图像模态	HFM*	83.44	76.57	84.15	80.18
	D&R Net*	84.02	77.97	83.42	80.60
	Att-BERT*	86.05	80.87	85.08	82.92
	InCrossMGs	86.10	81.38	84.36	82.84
	CMGCN*	86.54	—	—	82.73
	HKE	87.10	82.25	85.61	83.89
	Ours	88.07	82.66	87.45	84.99

注: *表示该结果为文献[5]中的实验结果

尽管上述使用多模态数据的基线模型从不同角度捕获了文本模态和图像模态的不一致性, 且性能也在不断改进, 但它们在对齐方面存在一定的不足. 相比之下, 本文模型使用多个自注意力机制来优化两种模态和两个模块的对齐, 并取得了一定的改进. 具体来说, 跨模态注意力对齐模块中的第 1 个自注意力网络用于增强冲突的文本 token 和图像 patch 的表示, 从而引导模型的注意力集中在冲突的文本 token 和图像 patch 对上. 预测模块中的第 2 个自注意力网络用于加强最终预测中高水平矛盾相关的信息, 使模型能够更专注于最终预测中更重要的矛盾信息.

4.4 消融实验

为了研究模型中每个独立组件对性能的影响, 我们进行了消融实验, 测试了去除某些组件后模型的性能. 具体来说, 我们研究了以下 3 种变体.

(1) 去除跨模态注意力对齐 (变体 1): 该模型去除了第 2 模块中的跨模态注意力网络.

(2) 去除图注意力网络 (变体 2): 该模型去除了第 3 模块中的图注意力网络.

(3) 去除预测模块中的自注意力 (变体 3): 该模型没有使用自注意力来对齐跨模态注意力模块和图注意力模块获取的信息.

结果如表 4 所示, 完整模型 (包括所有组件) 获得了最佳性能. 此外, 去除图注意力网络 (变体 2) 可能会导致性能大幅下降. 这表明有效捕获文本和图像之间的关系可能为讽刺检测提供了大量信息. 去除跨模态注意力网络 (变体 1) 和预测模块中的自注意力 (变体 3) 中的任意一个 (而不是两者), 也可能会导致一定程度的性能下降, 表明了文本和图像对齐优化的影响.

表 4 消融实验结果 (%)

变体	模型描述	准确率	*F*1
本文模型	完整模型	88.07	84.99
变体1	去除跨模态注意力对齐	83.48	81.51
变体2	去除图注意力网络	83.82	80.59
变体3	去除预测模块自注意力	86.14	81.87

我们还研究了 MCA 层和 GAT 层数量对模型性能的影响, 如图 3 和图 4 所示. 当 MCA 层为 5 层, GAT 层为 2 层时, 模型性能最佳.

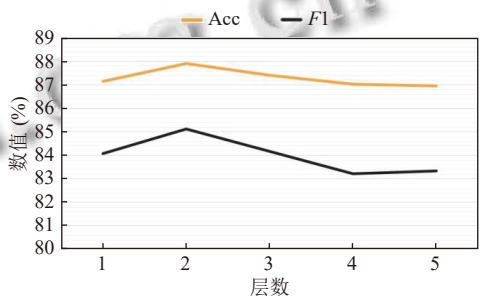

图 3 GAT 层数对模型性能的影响

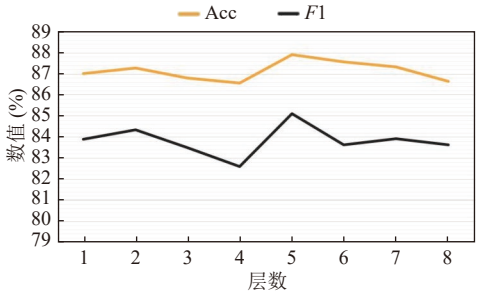

图 4 MCA 层数对模型性能的影响

4.5 案例分析

为了进一步解释模型工作的原因, 我们可视化了跨模态注意力模块和图注意力模块输出融合后的注意力分布, 以说明模型对多模态讽刺检测的有效性. 如图 5 所示, 在经过自注意力处理之前, 模型对所有图文对几乎一视同仁, 没有关注到图片中关于汽车凹陷的重要信息. 在经过自注意力处理之后, 模型能够根据信息的重要程度进行筛选, 最终能够关注到图片中的重要信息并与文本信息结合从而捕捉到不一致性.

(a) 处理前

(b) 处理后

图 5 经过自注意力处理前后的融合注意力分布对比

5 总结与展望

本文提出了一种多模态讽刺检测模型. 通过使用图注意力网络来捕捉文本和图像中不同元素之间的关系, 并使用注意力网络 (例如, 跨模态注意力, 自注意力) 在多个位置对齐文本和图像信息, 我们的模型能够有效地识别多模态数据中的对比, 进而检测出讽刺. 实验结果表明, 该模型在公共数据集上表现优于其他基准模型. 本研究的一个局限为我们只在一个公共数据集上评估了模型的有效性, 然而, 这是目前唯一可用于讽刺检测的公共数据集. 我们将在后面的研究中探索其他数据集. 另一个局限是我们没有测试包含外部知识的模型性能, 未来的工作将会完成这一部分的研究.

参考文献

1 Wang XY, Sun XW, Yang T, *et al*. Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data. Proceedings of the 1st International Workshop on Natural Language Processing Beyond Text. ACL, 2020. 19–29.

2 Liang B, Lou CW, Li X, *et al*. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. Proceedings of the 29th ACM International Conference on Multimedia. ACM, 2021. 4707–4715. [doi: 10.1145/3474085.3475190]

3 Liang B, Lou CW, Li X, *et al*. Multi-modal sarcasm detection via cross-modal graph convolutional network. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 1767–1777.

4 Pramanick S, Roy A, Patel Johns VM. Multimodal learning using optimal transport for sarcasm and humor detection. Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 3930–3940.

5 Liu H, Wang WY, Li HL. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 4995–5006. [doi: 10.18653/v1/2022.emnlp-main.333]

6 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1746–1751. [doi: 10.3115/v1/D14-1181]

7 Tay Y, Luu AT, Hui SC, *et al*. Reasoning with sarcasm by reading in-between. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 1010–1020. [doi: 10.18653/v1/P18-1093]

8 Schifanella R, de Juan P, Tetreault J, *et al*. Detecting sarcasm in multimodal social platforms. Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam: ACM, 2016. 1136–1145. [doi: 10.1145/2964284.2964321]

9 Cai YT, Cai HY, Wan XJ. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2506–2515. [doi: 10.18653/v1/p19-1239]

10 Xu N, Zeng ZX, Mao WJ. Reasoning with multimodal

sarcastic tweets via modeling cross-modality contrast and semantic association. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 3777–3786. [doi: 10.18653/v1/2020.acl-main.349]

11 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/n19-1423]

12 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

13 Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5-6): 602–610. [doi: 10.1016/j.neunet.2005.06.042]

14 Xiong T, Zhang PR, Zhu HB, *et al*. Sarcasm detection with self-matching networks and low-rank bilinear pooling. Proceedings of the 2019 World Wide Web Conference. San Francisco: ACM, 2019. 2115–2124.

15 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

16 Pan HL, Lin Z, Fu P, *et al*. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. Proceedings of the 2020 Findings of the Association for Computational Linguistics. ACL, 2020. 1383–1392. [doi: 10.18653/v1/2020.findings-emnlp.124]

(校对责编: 王欣欣)