郭圣濠, 蔡瑾, 晏峻峰. 基于图自注意网络的多模态讽刺检测[J]. 智能计算机与应用,2026,16(2):64-69. DOI:10.20169/j.issn.2095-2163.25073101

基于图自注意网络的多模态讽刺检测

郭圣濠[1,2], 蔡 瑾[1,2], 晏峻峰[1,2]
(1 湖南中医药大学 信息科学与工程学院, 长沙 410208; 2 湖南省智慧中医工程技术研究中心, 长沙 410208)

摘 要: 随着社交媒体的普及,讽刺检测成为了情感分析领域的一个重要任务。社交媒体文本的口语化和个性化特点使得传统的文本处理方法难以准确识别讽刺。为了提高检测的准确性,本文提出了一种基于图注意力网络(GAT)的多模态讽刺检测模型,该模型融合了图像和文本特征以捕捉讽刺的微妙细节。研究采用了 ResNet-50 来提取图像特征,该模型通过残差结构有效缓解深层网络训练困难,能够捕捉图像的高层语义特征。文本特征则通过 nghuyong/ernie-2.0-large-en 模型提取,该模型通过多任务自监督学习与知识融合,增强了模型的语义理解能力。进一步地,研究不是简单地连接不同模态的特征向量,而是使用 GAT 进行模态融合,利用特征节点计算余弦相似度构建动态矩阵,以同时考虑图像和文本的信息,提高跨模态任务的性能。最终,通过全连接层完成分类。研究在公开可用的多模式讽刺检测基准数据集上进行了实验。实验结果表明,所提的模型在讽刺检测任务上取得了较好的性能,证明了多模态特征融合和 GAT 应用的有效性。

关键词: 讽刺检测; 多模态学习; GAT
中图分类号: TP183 **文献标志码**: A **文章编号**: 2095-2163(2026)02-0064-06

Multimodal sarcasm detection based on graph self-attention networks

GUO Shenghao[1,2], CAI Jin[1,2], YAN Junfeng[1,2]
(1 School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China;
2 Hunan AI TCM Lab, Changsha 410208, China)

Abstract: With the widespread adoption of social media, sarcasm detection has become a critical task in sentiment analysis. The colloquial and personalized nature of social media texts poses challenges for traditional text processing methods in accurately identifying sarcasm. To enhance detection accuracy, this paper proposes a multimodal sarcasm detection model based on Graph Attention Networks (GAT), which integrates image and text features to capture the subtle nuances of sarcasm. The paper employs ResNet-50 to extract image features, leveraging its residual structure to mitigate training difficulties in deep networks and effectively capture high-level semantic features. Text features are extracted using the nghuyong/ernie-2.0-large-en model, which enhances semantic understanding through multi-task self-supervised learning and knowledge integration. Instead of simply concatenating feature vectors from different modalities, the paper utilizes GAT for modality fusion, constructing a dynamic adjacency matrix based on cosine similarity between feature nodes to simultaneously consider image and text information, thereby improving performance in cross-modal tasks. Finally, classification is performed through fully connected layers. Experiments conducted on a publicly available multimodal sarcasm detection benchmark dataset demonstrate that the proposed model achieves superior performance in sarcasm detection, validating the effectiveness of multimodal feature fusion and GAT application.

Key words: sarcasm detection; multimodal learning; Graph Attention Network(GAT)

0 引 言

根据剑桥词典,讽刺被定义为“使用明显与自己所说相反的言论来伤害某人的感情或以幽默的方式批评某事”。讽刺就像说一个命题的对立面一样,而这个命题本来是用真诚的话语表达的[1]。

讽刺识别的传统方法主要依靠人工构建特征模版和规则[2-3],但是人工构建特征模版和规则耗费专家的大量时间和精力,同时规则系统的可迁移性也差。

基金项目: 湖南省教育厅科学研究重点项目(23A312)。
作者简介: 郭圣濠(2000—),男,硕士研究生,主要研究方向:多模态融合,音乐情感分类; 晏峻峰(1965—),女,教授,博士生导师,主要研究方向:多模态融合,音乐情感分类。
通信作者: 蔡 瑾(1984—),女,博士研究生,主要研究方向:多模态融合,音乐情感分类。Email:caijin2014@qq.com。
收稿日期: 2025-07-31

随着讽刺检测成为情感分析的重要任务，也有学者将深度学习模型引入此研究中[4-6]。在当今社交媒体数据充斥的环境下，不再局限于自然语言处理领域。考虑到这一点，研究在文本检测讽刺的基础上，添加一个图像模态，即图片和其配文。现有的大多数多模态反讽检测模型中[7-10]，提取文本特征的使用的是 Bert[11] 及其衍生体、甚至是传统的 Word Embeddings，而本研究中采用的文本编码器为 nghuyong/ernie-2.0-large-en，该模型在 BERT 架构基础上进一步引入了多种知识增强机制，如实体、句法和语义信息的融合，具备更强的语言理解与情感建模能力，能够更准确地提取文本中的情感和语义特征；在图像特征提取方面，采用了经典的 ResNet-50 网络，该模型通过引入残差结构，有效缓解了深度网络中的梯度消失问题，同时在参数规模与性能之间取得了良好的平衡，因而在图像特征任务中表现出色。与更复杂的 DenseNet[12] 相比，ResNet-50 结构更为简介、训练更高效，适用于保持良好性能的同时降低计算成本的场景。

为了更有效地捕捉图像与文本之间的深层语义关系，本文在文献[13]工作的启发下，引入图神经网络对多模态特征进行融合，采用图注意力网络(Graph Attention Network，GAT)[14] 实现特征间的选择性信息传递与融合，从而提升了跨模态任务中的反讽识别能力。

本模型的创新点主要包括：

(1)多模态特征图建模与融合。通过建构图结构将图像与文本节点有机结合，并利用 GAT 实现信息的选择性融合，显著提升跨模态语义表达能力。

(2)基于 ERNIE 和 ResNet 的特征提取。分别利用了 ERNIE 2.0 large 和 ResNet-50 获取更具语音深度的文本与图像表示，较传统方法在表示能力上更优。

(3)引入图注意力机制。区别于 GCN 的固定加权方式，GAT 能够动态学习节点间的重要性，有效提升图中不同模态的融合效果。

综上可知，本文构建的 GAT 多模态反讽检测模型在保持结构简洁的同时，充分融合了图与文本模态的关键信息，具备更强的语义理解能力和情感判别性能，为多模态情感分析提供了更具实效的解决方案。

1　相关工作

1.1　传统文本处理方法

早期的讽刺检测研究主要依赖于传统的文本处理技术。例如，Khan 等学者[6] 提出了一种基于双向长短期记忆(Bi-LSTM)网络模型，该模型通过捕捉文本中的长期依赖关系来识别讽刺。然而，这些方法在处理口语化、个性化的社交媒体文本时，往往难以准确识别反讽信息。张庆林等学者[15] 提出了一种对抗学习[16] 框架，包含了 2 种互补的对抗学习方法，来克服现有讽刺识别方法的性能受到训练数据缺乏的影响。

1.2　多模态方法

多模态讽刺检测更多地关注多模态之间的信息融合、表示和信息相关性[17-19]。Das 等学者[7] 提出了一种并行深度 LSTM 框架(Parallel Deep LSTM)，用于更精确地检测和生成讽刺语言。该方法使用交叉激活函数，从而在学习过程中引入非线性交互，增强对反讽语言的建模能力。Cai 等学者[8] 提出了一种分层融合模型(Hierarchical Fusion Model)，文本通过 Bi-LSTM+Attention 进行建模，图像特征由 CNN 进行提取，最后通过融合网络整合 2 种模态，模型采用先文本后图像的方式建模多模态之间的语义对齐，处理图文不符的讽刺场景。Yue 等学者[9] 通过引入外部知识库(如 ConceptNet)，用于增强反讽检测过程中常识推理能力，提出了语文知识三元融合机制(Tri-Modal Knowledge Fusion)，结合上下文和图像间的语义关系，提高讽刺识别的准确率。Tian 等学者[10] 提出了一种创新的动态路由 Transformer 网络(DRT-Net)，模拟图文信息之间的非静态、语义依赖关系，该模型使用胶囊网络思想，构建了一种动态信息传递机制，能灵活应对不同模态的重要性变化，例如有些反讽推文图像比文本信息更关键的情况。

1.3　图神经网络在讽刺检测中的应用

图注意力网络(GAT)是一种图神经网络(GNN)的变体，在处理图结构数据方面具有独特的优势，已经用于多模态特征的融合。Li 等学者[20] 的工作表明，GAT 能够有效结合图像和文本特征，提高跨模态任务的性能。本研究在现有工作的基础上，提出了一种新的多模态讽刺检测模型，该模型结合了 ResNet-50 和 nghuyong/ernie-2.0-large-en 的特征提取能力，并通过 GAT 进行有效的特征融合，以提高讽刺检测的性能和准确性。

2　方法

图 1 显示了本文提出的图像和文本特征融合模型的体系结构。在该工作中，将一张图片和其配文输入模型。针对图片，将裁减的图片以及原图片输

入图像特征提取模块,进行预处理,得到一个特征向量 $\boldsymbol{v}_{\text{image}}$;文字由 nghuyong/ernie-2.0-large-en 预处理,得到一个代表输入文字的特征向量 $\boldsymbol{v}_{\text{text}}$。将 $\boldsymbol{v}_{\text{image}}$ 与 $\boldsymbol{v}_{\text{text}}$ 输入图卷积层,经过 2 层的 GAT 和 1 层全连接层得到一个融合特征向量 $\boldsymbol{v}_{\text{fusion}}$、大小为 5×1,代表了最初输入的图片和其配文。最后,$\boldsymbol{v}_{\text{fusion}}$ 进行全局平均池化和激活操作完成分类。

GAT-FIFT 通过处理多模态数据,利用图注意力网络(GAT)进行特征融合和分类,旨在检测社交媒体内容中的讽刺含义。整体框架包括特征提取、多模态特征融合、分类三个核心阶段,并通过数据增强、SMOTE 处理类不平衡、k 折交叉验证等技术提升性能,以下各小节将详细阐述每个模块。

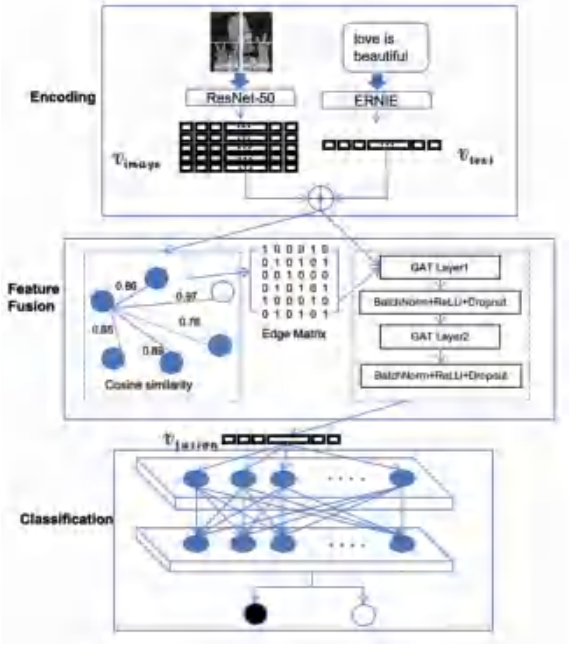

图 1 用于多模态讽刺检测的 GAT-FITF 总体架构

Fig. 1 General architecture of GAT-FITF for multimodal sarcasm detection

2.1 图像特征提取

首先将图像调整至 256×256,中心裁剪至 224×224,并进行归一化,生成输入张量。ResNet-50 模型(去除最后一层连接层)对输入张量进行前向传播,生成全局特征向量 $\boldsymbol{v}_{\text{global}} \in \mathbb{R}^{1\times 2048}$,通过全局平均池化捕捉图像的整体语义。

为增强局部特征表达,图像按 2×2 网格分割为 4 个子区域,每个子区域裁剪后调整至 224×224,分别输入 ResNet-50,生成 4 个局部特征向量 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4\}$,每个为 2 048 维。最终组合全局特征和 4 个局部特征,形成 5 个 2 048 维向量 $\boldsymbol{V} = \{\boldsymbol{v}_{\text{global}}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4\} \in \mathbb{R}^{5\times 2\,048}$。特征维度与文本特征对齐,便于后续多模态融合。

ResNet-50 通过深度卷积结构有效捕捉图像的全局和局部语义,适用于检测社交媒体中的讽刺视觉。网格分割策略增强了模型对图像细节的建模能力,确保特征的鲁棒性。

2.2 文本特征提取

文本特征提取模块旨在从社交媒体文本(如文本或评论)中提取语义表示,以支持多模态讽刺检测任务。本研究采用预训练的 ernie-2.0-large-en 模型[21],这是一种基于 Transformer 的增强表示模型,因其在处理复杂语义和上下文关系方面的优异性能而被选用。

对于每个文本数据,使用 ERNIE 的分词器(AutoTokenizer)将文本分词为子词序列,并在首尾添加[CLS]和[SEP]标记,得到序列:

$$S = [x_{\text{cls}}, x_1, x_2, x_3, \cdots, x_{\text{len}(S)}, x_{\text{sep}}] \tag{1}$$

其中,$x_i \in \mathbb{R}^V$ 表示词表大小,$\text{len}(S)$ 表示分词后的 token 数。设定文本最大长度为 $l = 512$,若 $\text{len}(S) \geqslant l - 2$,则截断为 $S' = [x_{\text{cls}}, x_1, \cdots, x_{l-2}, x_{\text{sep}}]$,若不足则以[PAD]补齐。

在分词之后,生成 inputs_ids 和 attention_mask,ernie-2.0-large-en 嵌入层将序列映射为词嵌入:

$$S_{\text{emb}} = \text{Emb}[x_{\text{cls}}, x_1, \cdots, x_{\text{sep}}] = [w_{\text{cls}}, w_1, \cdots, w_{\text{sep}}] \tag{2}$$

其中,$w_i \in \mathbb{R}^d, d = 1\,024$。嵌入序列输入 ERNIE 编码器,通过多头注意力机制生成上下文相关的隐层表示 $S_H = \text{ERNIE}[w_{\text{cls}}, w_1, \cdots, w_{\text{sep}}] = [w'_{\text{cls}}, w'_1, \cdots, w'_{\text{sep}}]$。取最后一层的[CLS]表示 $w'_{\text{cls}} \in \mathbb{R}^{1\times 1\,024}$ 作为文本特征。为与图像特征对齐,[CLS]表示通过线性层投影至 2 048 维,生成最终特征向量。批处理时,特征堆叠为 $R_{\text{emb}} \in \mathbb{R}^{N\times l\times d}$,其中 N 为批大小。

ERNIE-2.0-large-en 通过短语级预训练任务有效捕捉文本的讽刺语义,线性层确保多模态融合的维度一致性。

2.3 特征融合

本研究提出基于图注意力网络(GAT)的融合框架,通过动态图自适应地建模文本和图像特征的关系,显著提升检测性能。相较于传统融合方法(比如特征拼接或简单加权),GAT 通过注意力机制动态加权模态交互,特征适合处理讽刺表达中文本(如"sarcastic")与图像(如表情包)的非线性关联。

输入为每个样本的 6 个 2 048 维特征向量(1 个文本特征和 5 个图像特征),形状为 $\boldsymbol{X} \in \mathbb{R}^{N\times 6\times 2\,048}$,$N$ 为批大小,视为图的 6 个节点。特征首先通过线性

层预处理，映射至 1 024 维以降低计算复杂性：

$$\boldsymbol{X}' = \mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W}_p + \boldsymbol{b}_p) \tag{3}$$

其中，$\boldsymbol{W}_p \in \mathbb{R}^{2\,048\times 1\,024}$，$\boldsymbol{b}_p \in \mathbb{R}^{1\,024}$。

动态图结构基于余弦相似度构建，计算节点特征间的关联：

$$\mathrm{sim}(h_i, h_j) = \frac{h_i \cdot h_j}{\|h_i\| \cdot \|h_j\|} \tag{4}$$

仅保留相似度大于 0.8 的边，形成稀疏图；若无有效边，添加自环以确保连通性，生成边索引 $\mathbb{R}^{2\times E}$（E 为边数）。动态图避免了固定邻接矩阵的刚性假设，能够自适应地捕捉样本特定的模态关系，例如文本讽刺语气与图像视觉线索的强弱关联。

融合通过两层 GAT 实现，核心在于注意力机制动态加权邻居节点的贡献。第一层 GAT 有 5 个注意力头，计算注意力系数公式如下：

$$\alpha_{ij} = \mathrm{Softmax}_j(\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}h_i \parallel \boldsymbol{W}h_j])) \tag{5}$$

其中，$h_i \in \mathbb{R}^{1\,024}$；$\boldsymbol{W} \in \mathbb{R}^{1\,024\times 204}$；$\boldsymbol{a}$ 为注意力向量；“ ‖ ”表示拼接。节点特征更新为：

$$\alpha_{ij} = \mathrm{Softmax}_j h_i' = \mathrm{concat}\Big(\sum_{j\in N_i} \alpha_{ij}^{(k)} W^{(k)} h_j\Big)_{k=1}^{5} \tag{6}$$

输出 1 020 维（5×204）。第二层 GAT（将 1 020 维转换为 255 维，采用 5 头注意力机制）进一步精炼特征，生成 255 维表示（5×51）。每层 GAT 均结合批归一化和 dropout 防止过拟合，ReLU 激活增强非线性。GAT 输出重塑为 $\mathbb{R}^{N\times 1\,530}$（6×255），通过全连接层生成最终输出：

$$y = \boldsymbol{W}_2 \mathrm{LeakyReLU}(\boldsymbol{W}_1 h + b_1) + b_2 \tag{7}$$

其中，$\boldsymbol{W}_1 \in \mathbb{R}^{1\,530\times 256}$；$\boldsymbol{W}_2 \in \mathbb{R}^{256\times 1}$；$y$ 表示讽刺概率。

2.4　分类层

为实现多模态讽刺检测的二分类任务，本文设计了针对接受图注意力网络（GAT）融合后的多模态特征，生成讽刺/预测的结果。分类层采用 2 层全连接网络结构，接受 GAT 第二层输出的特征（被展平为 1 530 维向量），第一层全连接网络（FC1）将 1 530 维特征映射到 256 维，结合 LeakyReLU 激活函数（斜率 0.2），随后应用 dropout（概率 0.3）防止过拟合，增强非线性增强表达能力。最后，第二层全连接网络（FC2）将 256 维特征映射到单一输出，生成用于二分类的 logits。

3　实验

3.1　实验设置

3.1.1　数据集和评估指标

在实验中使用了公开可用的多模式讽刺检测基准数据集。该数据集由 Cai 等学者[8]创建，包含每条推特的图片和其对应的英文配文。此外，对数据集进行了处理，从图片中提取图像特征，从文本中提取文本特征。因此，每个样本都包括图像、文本特征。数据集的详细数据见表 1。

表 1　多模态讽刺检测基准数据集详细信息

Table 1　Details of the multimodal sarcasm detection benchmark dataset

统计内容	Training	Eval	Test
Sarcastic	5 006	411	407
Non-sarcastic	5 442	622	618
Total	10 508	1 033	1 025

本次研究中使用了准确率、精确率、召回率、$F1$ 分数作为评估指标。

3.1.2　基线模型

研究中将提出的 GAT-FITF 与多种已知的现有方法进行了比较，这里给出阐释分述如下。

（1）单图像模态。这些模型仅使用图像数据进行讽刺检测。ResNet 是一种基于 CNN 的残差连接图像分类器。ViT 是 Dosovitskiy 等学者提出的模型。将 Transformer 应用于图像分类。DenseNet 是一种深度卷积神经网络架构，具有密集的特征传递机制。

（2）单文本模态方法。这些模型仅基于用于讽刺检测的文本数据。Bi-LSTM 结构是一种递归神经网络，可以结合过去和未来的上下文信息来预测当前输出。同时，BERT 是一个基于 Transformer 的预训练语言模型。SKEP 是一个情感知识增强的预训练模型。

（3）多模态方法。这些模型使用图像和文本数据进行多模式讽刺检测。HFM 提出了一种用于多模式讽刺检测的分层多模式特征融合模型。VisualBERT 是一个预先训练的图像-文本模型，由一堆转换器层组成。最近的模型，如 D&R-Net、Res-Bert、Intra-Att、AttBert 基于注意力机制，而 InCrossMGs 和 CMGCN 基于图神经网络。

3.1.3　实验参数设置

使用 Focal Loss（$\gamma = 2.0, \alpha = 0.25$）作为损失函数，优化器为 Adam（学习率 0.000 05，$\beta_1 = 0.9$，$\beta_2 = 0.999$，权重衰减 0.01）。训练包含 L2 正则化（系数 0.1）和梯度裁剪最大范数（2.0）。学习率通过 ReduceLROnPlateau 调度器调整（衰减因子 0.5，耐心值 15，优化验证 $F1$ 分数）。批量大小为 256，最大训练轮数为 50 轮，采用 5 折交叉验证。训练使用加权随机采样（正样本权重为 4.0，负样本为 1.0），

并设置早停机制(耐心值为 10)。

3.2 实验结果和讨论

3.2.1 实验结果

在数据集上,将实验分为 3 组。GAT-FITF 模型与其他强大的现有模型之间的比较结果见表 2。

(1)文本模态方法:仅基于文本数据,BERT 就达到了最高的准确率(56.39%),SKEP 达到了最高的 $F1$ 分数(40.56%),这就表明在讽刺检测任务中有出色的性能表现。

(2)图像模态方法:ResNet 和 DenseNet 的准确率分别为 46.73%和 47.51%,非常接近。然而,与文本模态方法相比,图像模态方法表现不佳,这表明文本数据可能包含更有效的特征信息。如果研究中仅依靠图像数据进行讽刺检测,这将更具挑战性。

(3)多模态方法:将本文的 GAT-FITF 模型与为讽刺检测或其他多模式任务提出的一些强模型进行比较。分析可知,与单模态方法相比,大多数多模态方法的性能更好。然而,随着用于文本和图像处理的预训练模型的不断改进,一些单模态方法现在正超越以前的多模态方法的性能,如 BERT(83.85%)和 HFM(83.44%)间的性能差异。因此,本文使用相同的文本和图像预训练模型进行公平比较。GAT-FITF 模型与其他强大的现有模型之间的比较结果见表 2。表 2 中,带"*"的模型均基于 BERT 和 ResNet 模型。

表 2 GAT-FITF 模型与其他强大的现有模型之间的比较结果

Table 2 Comparison results between the GAT-FIFT model and other powerful existing models

Modality	Model	Acc	Pre	Rec	$F1$
Text	Bi-LSTM	41.90	56.66	48.42	47.53
	BERT	56.39	34.38	10.81	16.45
	SKEP	49.95	38.38	43.00	40.56
Image	ResNet	**46.73**	**34.79**	**39.07**	**36.81**
	Vit	**67.83**	**57.93**	**70.07**	**63.43**
	DenseNet	**47.51**	**35.73**	**40.29**	**37.88**
Multimodal	HFM	**83.44**	**76.57**	**84.15**	**80.18**
	D&R Net	**84.02**	**77.97**	**83.42**	**80.60**
	Res-Bert*	**84.80**	**77.80**	**84.15**	**80.85**
	Intra-Att*	**85.64**	**80.01**	**84.84**	**82.35**
	Att-Bert*	**86.05**	**78.63**	**83.31**	**80.90**
	InCrossMGs	**86.10**	**81.38**	**84.36**	**82.84**
	CMGCN*	**87.55**	**83.63**	**84.69**	**84.16**
	KnowleNet*	**88.87**	**88.59**	**84.18**	**86.33**
	GAT-FITF	**89.44**	**89.05**	**88.15**	**88.60**

从表 2 可以看出,本文模型在 4 个指标上均优于 KnowleNet、InCrossMGs 等近年来的多模态讽刺检测模型。

3.2.2 消融实验

为了进一步验证本文所提出的方法在讽刺识别领域的作用,对图注意力网络、余弦相似度构建动态图两个部分进行消融实验。实验结果见表 3。

表 3 消融实验结果

Table 3 Ablation experiment results

Configuration	Acc	Pre	Rec	$F1$
Baseline	89.44	**89.05**	**88.15**	**88.60**
w/o GAT	**77.03**	**65.21**	**618.00**	**73.40**
w/o Cosine Similarity for Dynamic Graph	**84.17**	**81.67**	**86.85**	**84.18**

表 3 的实验结果表明,去掉 GAT,也意味着去掉了计算余弦相似度构建动态图,直接进行文本和图像特征的拼接,以此来有效融合多模态,模型效果有一定下降,这是由于直接拼接特征忽视了模态间的复杂关系,无法捕捉模态之间的深层交互关系,并且拼接对文本和图片的特征一视同仁,无法根据任务需求动态调整各模态的贡献权重,而 GAT 通过注意力机制可以为不同模态分配不同的权重。

当去掉计算余弦相似度构建动态图模块之后,转而使用全连接图,模型的性能也有一定程度的降低。这可能是由于所有节点之间都存在边,无论是否相关,都会导致大量冗余边,模型可能会浪费计算资源去处理那些实际上不重要的关系,降低效率,并且所有节点都互相连接,噪声模态会直接影响所有其他节点。

从消融实验可知,在神经网络中加入图注意力网络和计算余弦相似度构建动态矩阵模块,能够提高多模态讽刺识别的性能。

3.2.3 图注意力神经网络参数分析

本文还分析了不同的图注意力神经网络层数对多模态讽刺识别的影响,实验结果见表 4。

表 4 不同的 GAT 层数对实验影响结果

Table 4 Experimental results of different GAT layers

GAT_num	Acc	Pre	Rec	$F1$
1	84.35	86.64	79.72	83.03
3	84.38	85.58	81.33	83.40
4	83.88	80.73	87.19	83.59
5	84.00	83.26	84.07	83.66
6	79.19	76.65	80.80	78.67
7	80.07	84.27	71.46	77.34
2	89.44	89.05	88.15	88.60

由表 4 可知,在 GAT 层的层数为 2 的时候,性

能最佳。

4　结束语

本文提出了一种基于 GAT 的多模态讽刺识别模型来提高在日常社交中的讽刺识别精度。通过 ResNet-50 和 nghuyong/ernie-2.0-large-en 的特征提取能力来对多模态数据进行提取特征，然后通过 GAT 模块和计算余弦相似度构建动态图来进行多模态特征的有效融合，通过与其他多模态讽刺检测的强大模型对比、消融实验、GAT 层数实验来证明研究提出模型的性能更优。本模型通过使用预训练的 ResNet-50 和 nghuyong/ernie-2.0-large-en 模型，显著减少了计算量和训练时间。在今后的研究中，可以尝试使用对预训练网络的微调，以此尝试是否可以进一步提高多模态讽刺的精度。

参考文献

[1] BOUAZIZI M, OHTSUKI T. Sarcasm over time and across platforms: Does the way we express sarcasm change? [J]. IEEE Access, 2022, 10: 55958-55987.

[2] KREUZ R, CAUCCI G. Lexical influences on the perception of sarcasm[C]//Proceedings of the Workshop on Computational Approaches to Figurative Language (FigLanguages′07). New York: ACM, 2007: 1-4.

[3] CARVALHO P, SARMENTO L, SILVA M J, et al. Clues for detecting irony in user-generated contents: oh...!! it′s" so easy"; -[C]//Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. New York: ACM, 2009: 53-56.

[4] BAMMAN D, SMITH N. Contextualized sarcasm detection on twitter[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2015, 9(1): 574-577.

[5] 刘龙飞，杨亮，张绍武，等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报，2015,29(6):159-165.

[6] KHAN A, MAJUMDAR D, MONDAL B, et al. A deep learning approach to sarcasm detection from composite textual data[J]. INFOCOMP Journal of Computer Science, 2022, 21(2):1-9.

[7] DAS S, GHOSH S, KOLYA A K, et al. Unparalleled sarcasm: A framework of parallel deep LSTMs with cross activation functions towards detection and generation of sarcastic statements[J]. Language Resources and Evaluation, 2023, 57(2): 765-802.

[8] CAI Yitao, CAI Huiyu, WAN Xiaojun. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL,2019: 2506-2515.

[9] YUE Tan, MAO Rui, WANG Heng, et al. KnowleNet: Knowledge fusion network for multimodal sarcasm detection[J]. Information Fusion, 2023, 100: 101921.

[10] TIAN Yuan, XU Nan, ZHANG Ruike, et al. Dynamic routing transformer network for multimodal sarcasm detection[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, 2023: 2468-2480.

[11] DEVLIN J, CHANG M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). ACL, 2019: 4171-4186.

[12] TIAN Hao, GAO Can, XIAO Xinyan, et al. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis[J]. arXiv preprint arXiv,2005.05635, 2020.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ: IEEE, 2016: 770-778.

[14] MOHAN A, NAIR A M, JAYAKUMAR B, et al. Sarcasm detection using bidirectional encoder representations from transformers and graph convolutional networks[J]. Procedia Computer Science, 2023, 218: 93-102.

[15] 张庆林，杜嘉晨，徐睿峰. 基于对抗学习的讽刺识别研究[J]. 北京大学学报(自然科学版)，2019，55(1)：29-36.

[16] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv, 1312.6199, 2013.

[17] XIA Yingjie, ZHANG Luming, LIU Zhenguang, et al. Weakly supervised multimodal kernel for categorizing aerial photographs[J]. IEEE Transactions on Image Processing, 2016, 26(8): 3748-3758.

[18] LU Jiasen, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. arXiv preprint arXiv, 1908.02265,2019.

[19] YAO Ziyan, LIN Fei, CHAI Sheng, et al. Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis[C]//Proceedings of 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE). Piscataway, NJ: IEEE, 2024: 1217-1223.

[20] LI Jiang, WANG Xiaoping, LV Guoqing, et al. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation[J]. Neurocomputing, 2023, 550: 126427.

[21] SUN Yu, WANG Shuohuan, LI Yukun, et al. ERNIE 2.0: A continual pre-training framework for language understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8968-8975.