论文Mimicking the brain’s cognition of sarcasm from multidisciplines for Twitter sarcasm detection复现

摘要

讽刺是一种表达轻蔑或嘲笑的复杂结构，通常用于表达与字面意思相反的意思，揭示社会问题或批评特定现象。单一模态的讽刺检测可能受到上下文缺失或不完整信息的影响，而多模态方法通过整合信息，有助于减少误解。本文通过利用视觉大模型来生成图像的文本描述，从而补充和增强早期的语义分析。设计特征融合模块融合不同模态的信息，并引入对比学习方法对齐视觉和文本的特征。在相同公开数据集上的实验结果显示，本文提出方法可以有效提高多模态讽刺检测任务的各项评价指标,检测精度提高了超过9%。

关键词：讽刺检测；多模态；情感计算；大模型

1 引言

讽刺是一种富有表现力的语言艺术，通常用于表达与字面意思相反的意思，揭示社会问题或批评特定现象。随着社交媒体的蓬勃发展，用户不仅通过文字传递讽刺，还结合了图像、视频、音频等多种媒体形式。这种现象促使了“多模态讽刺检测”的研究，旨在通过整合不同模态的信号，提升对讽刺内容的理解与判别。

多模态数据的特点是融合了多种信息来源，例如：文本：讽刺性评论或标题。图像：配图、表情包、漫画等视觉元素。视频：短视频内容中的表情、语调和动作。音频：语音语调、停顿以及其他声调变化。单一模态在讽刺检测中可能无法充分捕捉表达意图，因此，多模态方法的引入提供了更为全面的视角，挑战同时也带来了机遇。

多模态讽刺检测将文本、图像结合，有助于更准确地理解信息。讽刺的意图常常依赖于复杂的上下文，以及不同模态间的相互作用。在多模态框架下，机器能够分析多种信号，提升理解的深度和广度。单一模态的讽刺检测可能受到上下文缺失或不完整信息的影响，而多模态方法通过整合信息，有助于减少误解。例如，某个文本可能在没有上下文的情况下难以判断，但如果结合了相关图像或视频片段，机器就能更加准确地识别讽刺意图。在新时代的社交平台上，信息的传播往往是跨媒体的。通过多模态讽刺检测，可以研究不同渠道中讽刺的变体及其传播方式，从而帮助理解信息的流动与变迁。讽刺常常反映了特定社群对社会、文化和政治的态度。通过多模态分析，研究者可以获得对公众情感和意见的更深刻理解，从而为舆情监测和公共政策的制定提供参考。随着社交媒体的不断发展，研究者继续深化多模态讽刺检测的技术，将有助于提升信息理解的准确度，促进社会文化的全面认知。

在当前的讽刺检测任务中，虽然自然语言处理（NLP）技术已经取得了一定的进展，但往往忽视了图像数据所蕴含的丰富语义信息。许多讽刺表达不仅依赖于文本元素，还深受图像上下文的影响。例如，在社交媒体上，许多讽刺性的评论往往是与特定的图片、漫画或表情包结合的。这些视觉元素承载了许多隐含的情感、背景和讽刺意图，单靠文本分析难以全面捕捉。

为了应对上述挑战，本文提出了一种新方法，通过利用视觉大模型来生成图像的文本描述，从而补充和增强早期的语义分析。设计特征融合模块融合不同模态的信息，并引入对比学习方法对齐视觉和文本的特征。千问视觉语言大模型是自然语言处理(NLP)和计算机视觉(CV)领域的一项前沿技术，它结合了图像和文本两种类型的数据处理能力，能够在多个任务上表现出色，例如图像理解、文本生成和跨模态转换等，本文使用Qwen2-VL生成图像的文本描述。本研究实现了以下几个关键目标：丰富的语义补充:视觉大模型通过分析图像特征、模式和上下文，提取出丰富的语义信息。这些信息被转化为文本描述，能够捕捉到图像深层次的含义，补充了仅依赖文本分析的局限性。多模态融合:通过将图像、生成的文本描述与原始文本应用于多模态讽刺检测模型，我们能够更好地整合不同模态的信息。这样的融合不仅增加了信息的多样性，还增强了模型对讽刺意图的判别能力。上下文理解:生成文本描述的方法使得模型更好地理解图像与文本之间的上下文关系。这种上下文理解极大提高了讽刺检测的准确率，因为不少讽刺语句的有效性在于其与图像内容的相互影响。

本文的贡献如下：

1. 使用视觉语言大模型丰富图像的语义信息；

2. 引入对比学习方法对齐模态间的语义信息，设计多模态融合网络，构建多模态讽刺检测模型。

2 相关工作

文献 [3]提出了一种基于情感感知的分层融合网络用于多模态讽刺检测。该方法先通过预训练模型提取图像、文本和属性特征，在融合层中，利用情感感知层将各模态的情感向量与原始特征融合，得到情感感知特征；再通过多模态交互层的跨模态Transformer学习模态间依赖关系；然后在记忆融合层中计算各模态的记忆向量，并进行特征融合得到最终向量；同时，使用文本-图像对比损失来增强模态间语义对齐，最后通过讽刺分类层预测讽刺标签。

文献 [8]探讨了利用多学科知识（神经解剖学和神经心理学）来改进Twitter讽刺检测的方法，提出了一种名为的多模态、多交互和多层次神经网络模型,由提取、交互与合作、集成和认知四个部分组成，从下到上与大脑对讽刺的认知过程一致。探索模型在多模态情感分析和多模态情绪识别任务中的性能，验证其泛化能力和在相关任务中的优越性。

现有的讽刺检测方法的图像编码器倾向于将相似的图像编码为相似的向量，并且由于GAT层的累积和非相邻节点的缺乏表示引起的负相关性，在图级特征提取中引入了噪声。为了解决这些局限性，文献 [7]提出了一种双层自适应不一致性增强模型（DAIE），以提取文本和图像在标记和图形级别上的不一致性。在表征层面，使用基于补丁的重建图像来支持表征层面的对比学习，以捕捉图像的共同和特定特征，从而放大文本和图像之间的不一致性。在图像级别，引入自适应图对比学习，并结合负对相似性权重，以细化模型文本和视觉图节点的特征表示，同时增强相邻节点之间的信息交换。

一些对讽刺检测的研究提供了对各个方面不一致的粗粒度检测，忽略了情感和句法方面不连贯的多方面特征。文献 [9]从对讽刺特征的更细粒度分析入手，提出了一种名为MDSAN（多维深层语义对齐网络）的新模型。MDSAN模型由两个不一致性捕获模块（语义提取模块（SEM）和情绪倒置检测模块（EID））和一个深度跨模态蒸馏和对齐模块（DCIE）组成。SEM模块整合了先验知识，并利用多头机制从多层中挖掘不连贯的文本语义信息。EID模块从Kantorovich问题开始，使用Sinkhorn算法最小化情感空间中文本和图形分布的成本流，量化文本图像对的情感反转程度，以协助讽刺检测。

多模态讽刺检测任务面临着两个潜在的挑战。首先，利用单独预训练的单峰模型提取视觉和文本特征通常缺乏有效多模态数据集成所需的基本对齐能力。其次，视觉和语言之间有害的模态差距使得仅通过不同的跨模态融合技术全面整合多模态信息变得具有挑战性。文献 [6]我们提出了一种多模式相互学习（MuMu）网络来解决这些问题。使用大规模对比语言图像预训练模型中的图像和文本编码器初始化MuMu网络，以增强底层图像与文本的对应关系。此外，为了提高融合过程中捕获跨模态不一致的能力，设计了一种对齐-融合协作机制，在融合前对齐不同的模态，并在融合后通过互学习增强两种模态之间的协作建模能力。

文献 [5]指出以前的工作主要是设计复杂的网络结构来融合图像文本模态特征进行分类。然而这种复杂的结构可能会对域内数据进行过拟合，从而降低非分布（OOD）场景中的性能。此外，现有的方法通常没有充分利用跨模态特征，限制了它们在域内数据集上的性能。因此，为了构建更可靠的多模态讽刺检测模型，作者提出了一种生成式多模态讽刺模型，该模型由一个设计的指令模板和一个基于大型语言模型的演示检索模块组成。

文献 [4]指出之前的检测不一致方法主要集中在语义层面，往往忽视了更具体的讽刺不一致形式。特别是讽刺不一致，包括事实不一致、情感不一致和组合不一致。因此，作者从新颖的角度提出了一个事实-情感不一致组合网络，通过探索多模态事实差异、情感不一致和组合融合来绘制多模态讽刺关系。设计了一个动态连接组件，通过图注意力和掩码路由矩阵计算动态路由概率权重，选择最合适的图像-文本对来捕捉图像和文本之间的事实不一致性。然后使用外部情感知识检索文本标记和图像对象之间的情感关系，重建跨模态图矩阵中的边缘权重，以捕捉情感不一致。此外，引入了一个组合不一致融合层和跨模态对比损失来融合事实不一致和情感不一致，以进一步增强不一致表征。

文献 [3]引入了一种分层融合模型，整合了情感信息以增强多模态讽刺检测。在图像模态中使用属性对象匹配，将其视为辅助属性模态。然后从每种模态中提取情感数据，并将其组合在一起，以在模态中实现更全面的表示。此外，使用交叉模态变换器来表征模态间不一致性的关系。我们还实现了一种情感感知的图像文本对比丢失机制，以更好地同步图像和文本的语义。通过强化这些对齐，模型能够更好地理解不协调的关系。

上面介绍的讽刺检测研究工作从多模态讽刺检测的数据特征提取、特征对齐与融合角度出发，致力于提高模型对图像和文本的不一致性感知，但往往忽略了图像包含的丰富的语义信息。目前的算法对于自然语言的能力得到充分释放，而对图像信息的挖掘是不充分的。随着大模型技术的发展，人们提出了视觉语言大模型（如Qwen-VL [1]、LLaMA [2]等），增强机器对图像的理解能力。本文借助于视觉语言大模型对图像的强大理解能力，对图像生成对应的文本描述，用文本描述图像的包含的丰富场景和表达的情感，作为辅助模态。结合图像、帖子文本、图像的描述文本这三种模态数据，通过引入对比学习方法对齐图像模态和文本模态的语义信息，设计基于注意力的特征融合模块，构建模态增强的多模态讽刺检测方法。

3 本文方法

3.1 符号定义

给定一组N条多模态推文$\mathcal{D}=\{\langle(\mathbf{x}_1^{\text{img}},\mathbf{x}_1^{\text{twt}}),\mathbf{y}_1\rangle,\ldots,\langle(\mathbf{x}_N^{\text{img}},\mathbf{x}_N^{\text{twt}}),\mathbf{y}_N\rangle\}$，每个推文包含文本-图像 $(\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}})$ 和对应的标签$\mathbf{y}_i\in\{0,1\}$，目标是设计一个模型 $\mathcal{F}(\cdot)$获取第 i 条推文的概率分布 $\mathcal{D}_i$。神经心理学研究指出，感知讽刺需要多种模态。因此，我们还使用从给定图像生成的图像描述和图像中的文本作为额外的模态，因为它们可以提供重要信息和额外的图像属性。

$$
\begin{aligned}
&p(\mathcal{D}_i)=p(\mathbf{y}_i\mid\langle\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}}\rangle)\\
&=p\left(\mathbf{y}_i|<\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}},\mathbf{x}_i^{\text{cap}}|\mathbf{x}_i^{\text{img}},\mathbf{z}>\right)\\
&\mathcal{F}(\mathcal{D}_i)=\mathcal{F}\left(\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}}\right)\\
&=\mathcal{F}(\langle\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}},\mathbf{x}_i^{\text{cap}}\rangle\mid\mathbf{z})\\
&=\mathcal{F}\left(<\mathbf{x}_i^{\text{img}},\mathbf{x}_i^{\text{twt}},\mathcal{F}^{\text{cap}}\left(\mathbf{x}_i^{\text{img}}\right)>|\mathbf{z}\right)
\end{aligned}
$$

3.2 本文方法概述

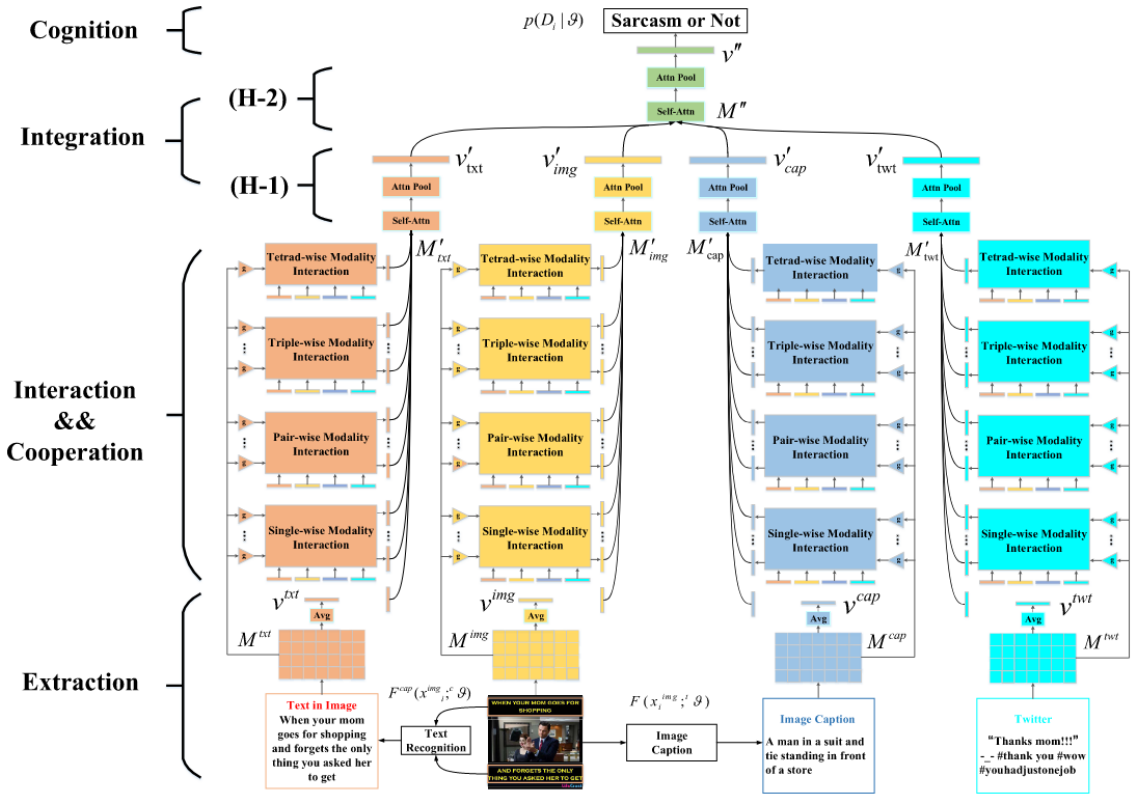

图 1. 论文模型结构示意图

论文提出的M_3N_2模型结构旨在模拟大脑对讽刺的认知过程，从多学科角度进行设计，模型由四个部分组成，从下到上依次为提取（Extraction）、交互与合作（Interaction And Cooperation）、集成（Integration）和认知（Cognition），与大脑对讽刺的认知过程相对应。

提取（Extraction）：

将图像调整为331×331大小，并划分为121个11×11区域。对每个区域应用预训练的NasNet获取区域图像表示，形成121×dimg（dimg初始为4032，经三层线性投影和ReLu激活函数后降为512）的图像记忆矩阵M^{img}。通过平均池化得到图像指导向量v^{img}。

使用“Show, Attend, and Tell (SAT)”模型生成图像说明X^{cap}，将每个单词嵌入向量后用BiLSTM编码。形成k×dcp（dcp = 512）的图像标题记忆矩阵M^{cap}，再平均得到指导向量v^{cap}。

利用ASTER识别图像中的文本，得到m×dtt（dtt = 512）的记忆矩阵M^{txt}。通过平均得到指导向量v^{txt}。

将Twitter视为单词序列，用双向LSTM编码，初始化时用图像指导向量v^{img}初始化BiLSTM的初始状态。生成的双向隐藏状态组成n×dwt（dwt = 512）的记忆矩阵M^{twt}，经平均池化得到推特指导向量v^{twt}。

交互与合作（Interaction And Cooperation）：

1. 单模态交互（Singlewise Modality Interaction）：

确保每个模态能从其他模态获取补充信息，如Twitter模态可关注图像模态中的相关内容。对于每个交互，先对一个模态的记忆矩阵（如M^{img}）应用门机制减少冗余，然后用GA从该矩阵中提取与另一个模态（如Twitter模态的v^{twt}）相关的信息v_0^{twt}。接着通过多跳过程（MH）从图像模态中获取多视角信息，最后用注意力池化合成向量（如$\tilde{v}^{twt}$）。共有$C_4^1 = 4$种这样的交互，产生四个输出向量。

2. 成对模态交互（Pairwise Modality Interaction）：

期望任意两个模态能从特定模态中找到互补或辅助信息，如Twitter和图像模态可从图像标题中获取相关信息。先融合两个模态的指导向量（如$v^{twt\&img} = v^{twt} + v^{img}$），对特定模态的记忆矩阵（如$M^{cap}$）应用门机制去噪。用GA关注图像标题中对两个模态有用的部分，再通过多跳找到相关部分，最后用注意力池化合成向量（如$\tilde{v}^{twt\&img}$）。共有$C_4^2 = 6$种交互，产生六个输出向量。

3. 三元模态交互（Triplewise Modality Interaction）：

从一个模态中同时提取与三个模态相关的信息，如从图像标题中为图像、Twitter和图像中的文本三个模态提取相关知识。先融合三个模态的指导向量（如$v^{twt\&img\&txt} = v^{twt} + v^{img} + v^{txt}$），对特定模态的记忆矩阵（如$M^{cap}$）应用门机制去噪。用GA关注图像标题中与三个模态相关的部分，再通过多跳找到相关部分，最后用注意力池化合成向量（如$\tilde{v}^{twt\&img\&txt}$）。共有$C_4^3 = 4$种交互，产生四个输出向量。

4. 四元模态交互（Tetradwise Modality Interaction）：

融合四个模态的指导向量与一个记忆矩阵交互，产生一个向量（如$\tilde{v}^{img\&cap\&txt\&twt}$），交互方式与前面类似，只是输入为四个模态的指导向量之和。

集成（Integration）：

1. 层次 - 1（Hierarchy - 1，H - 1）：

对每个模态的新矩阵（如M'_{mode}）应用自注意力（SA）后再进行注意力池化，将其合成并浓缩为向量（如v'_{mode}），代表该模态的一阶理解。四个模态产生四个向量，组成新矩阵M''。

2. 层次 - 2（Hierarchy - 2，H - 2）：

对M''再次应用自注意力（SA）和注意力池化，将所有模态信息集成到一个最终向量v''，

用于最终的讽刺认知。

认知（Cognition）：

在v''上应用两个线性层，分别由ReLu和sigmoid激活，以区分是否为讽刺。使用交叉熵作为损失函数进行训练，采用F_1 - score和准确率作为评估指标。

4 复现细节

4.1 与已有开源代码对比

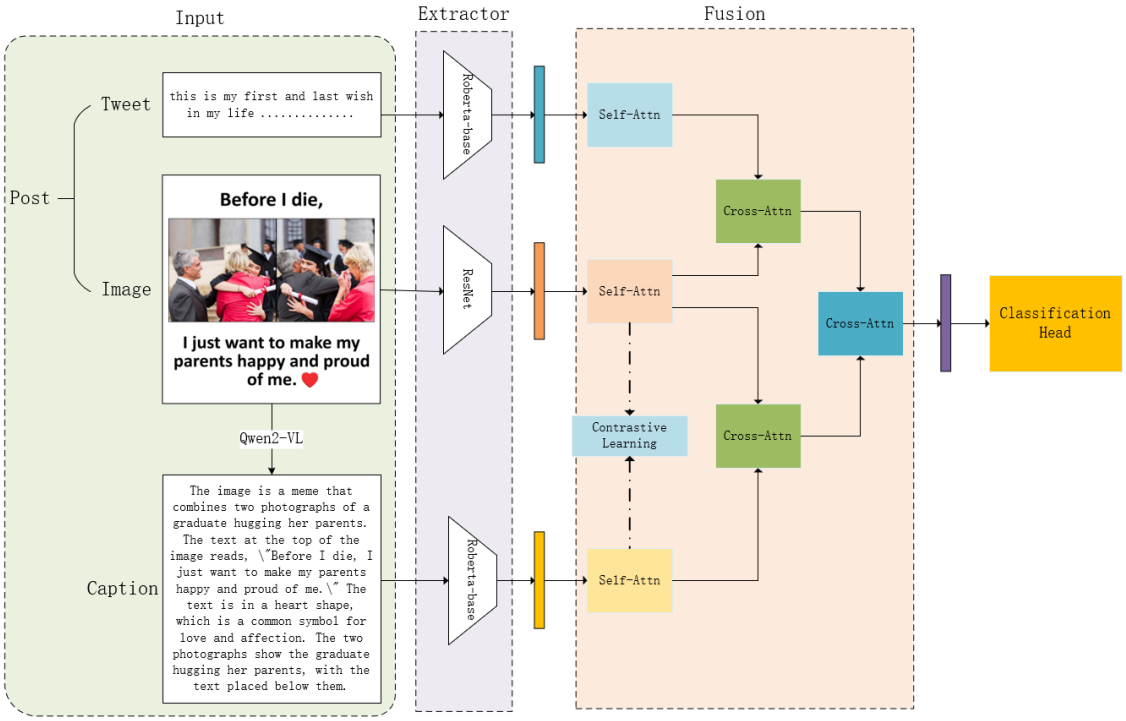

图 2. 模型改进结构示意图

论文未开源源代码，本复现未参考相关源代码。

复现中，由于论文使用的一些方法老旧，效果不好，故替换为较新的方法。并通过使用Qwen2-VL视觉语言模型生成图像对应的Caption文本，设计新颖的特征融合模块，引入对比学习方法对齐图像和文本的语义信息。

提出的模型结构如图2所示，原始数据集由图像和Tweet文本构成，本文使用Qwen2-VL视觉语言模型生成图像对应的Caption文本。在特征提取阶段，分别使用Roberta-base和ResNet提取文本和图像的特征表示，得到Tweet、Image、Caption的特征向量。在特征融合阶段，首先对不同模态的特征向量做自注意力计算，然后Tweet和Image、Caption和Image分别做交叉注意力计算，对得到的两个向量结果再做交叉注意力，得到最后的融合后的特征。接上分类头，进行讽刺检测的预测。

为了对齐Image和Caption之间的语义信息，引入对比损失函数，

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\sin(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_{j=1}^{N}\exp(\sin(\mathbf{z}_i, \mathbf{z}_j)/\tau)}$$

其中， z_i表示第i个样本的特征表示， z_i^+表示第i个样本的正样本表示。 $\sin(\cdot,\cdot)$表示余弦相似度。

$$\sin(\mathbf{z}_1, \mathbf{z}_2) = \frac{\mathbf{z}_1^\top \mathbf{z}_2}{||\mathbf{z}_1||||\mathbf{z}_2||}$$

交叉熵损失如下：

$$\mathcal{L}_{bce} = -w_n\left[y_n \cdot \log x_n + (1-y_n)\cdot\log(1-x_n)\right]$$

最终的损失函数如下：

$$\mathcal{L} = \lambda\mathcal{L}_{\text{InfoNCE}} + (1-\lambda)\mathcal{L}_{bce}$$

4.2 创新点

使用Qwen2-VL视觉语言模型生成图像对应的Caption文本，设计新颖的特征融合模块，引入对比学习方法对齐图像和文本的语义信息。

5 实验结果分析

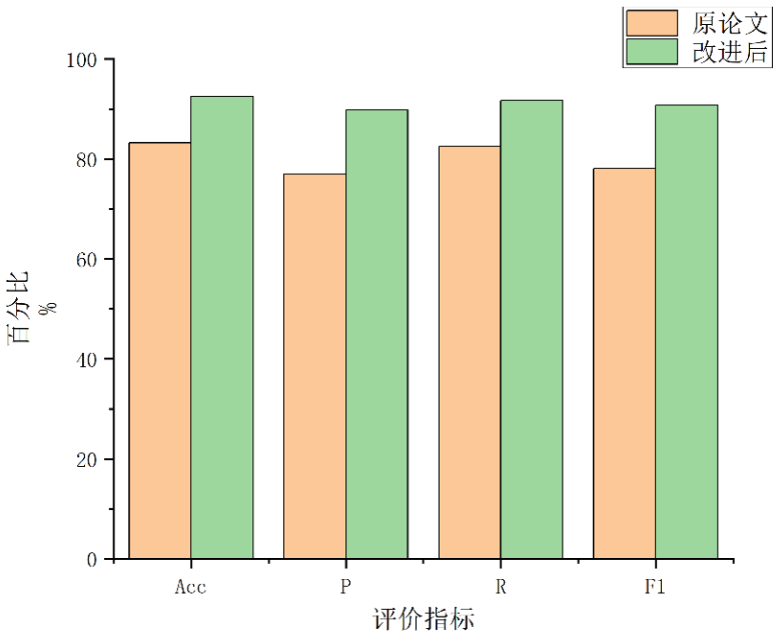

图 3. 实验结果示意

	Models	Acc	P	R	F_1
	Random	50.27	40.05	50.57	44.70
Single-Modal	[7]	77.50	67.25	84.78	75.00
	[29]	80.57	75.55	75.70	75.63
	[48]	80.90	76.46	75.18	75.82
	[8]	79.74	71.99	80.40	75.96
	[28]	80.94	73.72	81.02	77.20
Multi-Modal	[34]	80.41	72.40	82.06	76.93
	[9]	81.62	74.21	82.64	78.10
	M_3N_2	83.23	77.02	82.48	79.66
	Ours	92.56	89.79	91.76	90.76

表 1. 讽刺检测评价指标对比

实验结果如表1所示，实验证明本文提出的方法比其他对比方法的讽刺检测指标有显著提高。可以证明本文提出方法的有效性，并说明利用视觉语言大模型对图像的理解能力可以提高模型对数据的挖掘能力，从而提高讽刺检测任务的效果。

6 总结与展望

本文对论文Mimicking the brain’s cognition of sarcasm from multidisciplines for Twitter sarcasm detection做了复现，使用Qwen2-VL视觉语言模型生成图像对应的Caption文本，设计新颖的特征融合模块，引入对比学习方法对齐图像和文本的语义信息，改进了算法的检测效果。之后会进一步做更充分的实验验证，撰写学术论文。

参考文献

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama interface for vision tasks. *arXiv preprint arXiv:2403.00522*, 2024.

[3] Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Cheng Liu, and Dazhi Jiang. Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. *Information Fusion*, 108:102353, 2024.

[4] Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion*, 104:102203, 2024.

[5] Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceed-*

ings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1732–1742, 2024.

[6] Jie Wang, Yan Yang, Yongquan Jiang, Minbo Ma, Zhuyang Xie, and Tianrui Li. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103:102132, 2024.

[7] Qiaofeng Wu, Wenlong Fang, Weiyu Zhong, Fenghuan Li, Yun Xue, and Bo Chen. Dual-level adaptive incongruity-enhanced model for multimodal sarcasm detection. *Neurocomputing*, 612:128689, 2025.

[8] Fanglong Yao, Xian Sun, Hongfeng Yu, Wenkai Zhang, Wei Liang, and Kun Fu. Mimicking the brain’ s cognition of sarcasm from multidisciplines for twitter sarcasm detection. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):228–242, 2021.

[9] Bengong Yu, Haoyu Wang, and Zhonghao Xi. Multifaceted and deep semantic alignment network for multimodal sarcasm detection. *Knowledge-Based Systems*, 301:112298, 2024.