第 13 卷 第 5 期
2024 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 13 No. 5
Sep. 2024

引文格式：
夏冰, 杨瑞楠, 董玉, 等. 一种多模态隐喻数据集的构建和验证方法 [J]. 集成技术, 2024, 13(5): 64-73.
Xia B, Yang RN, Dong Y, et al. A method for constructing and validating a multimodal metaphor dataset [J]. Journal of Integration Technology, 2024, 13(5): 64-73.

一种多模态隐喻数据集的构建和验证方法

夏 冰* 杨瑞楠 董 玉 楚世豪 唐崇俊 葛云翔 尹家斌

（中原工学院 郑州 450007）

摘 要 隐喻的目的是启发理解、说服他人。目前，隐喻呈现文本、图像、视频等多模态融合的趋势，因此，识别多模态信息中蕴含的隐喻语义对互联网内容安全具有研究价值。由于缺乏多模态隐喻数据集，难以建立研究模型，因此，当前学者更关注基于文本的隐喻检测。针对这一不足，作者首先从图像-文本、隐喻出现、情感表达和作者意图等角度构建新型多模态隐喻数据集；其次，对数据集的标注者进行 Kappa 分数计算；最后，借助预训练模型和注意力机制融合图像属性特征、图像实体对象特征和文本特征，构建多模态隐喻检测模型，验证多模态数据集的质量和价值。实验结果表明：具有情感和意图表达的隐喻数据集可提升隐喻模型检测效果，多模态信息间相互关系有助于隐喻的理解。

关键词 内容安全；多模态隐喻检测；外部知识；多模态数据集；注意力机制
中图分类号 TP181 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20240124001

A Method for Constructing and Validating a Multimodal Metaphor Dataset

XIA Bing* YANG Ruinan DONG Yu CHU Shihao TANG Chongjun GE Yunxiang YIN Jiabin

(*Zhongyuan University of Technology*, *Zhengzhou* 450007, *China*)

*Corresponding Author: *xiabing@zut.edu.cn*

Abstract Metaphor has the purpose of inspiring understanding and persuading others. Currently, metaphor presents the trend of multimodal integration of text, images, and videos. Therefore, identifying the metaphorical semantics contained in multimodal contents has research value for Internet content security. Due to the lack of multimodal metaphor datasets, it is difficult to establish research models. Therefore, current scholars pay more attention to text-based metaphor detection. To overcome this shortcoming, the paper first generates a new multimodal metaphor dataset from the perspectives of image-text, metaphor appearance, emotion expression, and author intention. Then, Kappa scores were used to assess the consistency among the

收稿日期：2024-01-24 **修回日期**：2024-04-28
基金项目：河南省科技攻关项目（232102211088）；河南省哲学社会科学规划项目（2022BXW018）
作者简介：夏冰（通讯作者），博士，副教授，研究方向为网络安全和网络舆情，E-mail：xiabing@zut.edu.cn；杨瑞楠，硕士研究生，研究方向为网络安全和多模态隐喻；董玉，硕士研究生，研究方向为网络安全；楚世豪，硕士研究生，研究方向为网络安全；唐崇俊，硕士研究生，研究方向为网络安全；葛云翔，硕士研究生，研究方向为网络安全；尹家斌，硕士研究生，研究方向为网络安全。

annotators of the dataset. Finally, a multimodal metaphor detection model is constructed to verify the quality and value of the multimodal data set by combining image attribute features, image entity features, and text features with the help of a pre-training model and attention mechanism. The experimental results show that the metaphor dataset with emotion and intention can improve the effectiveness of metaphor model detection, and confirm that the interrelationship of multimodal information is helpful for understanding metaphor.

Keywords　content security; multimodal metaphor detection; external knowledge; multi-modal dataset; attention mechanism

Funding　This work is supported by Henan Science and Technology Research Project (232102211088), Henan Philosophy and Social Sciences Planning Project (2022BXW018)

1　引　言

多媒体信息的迅速发展使得互联网上涌现大量多模态数据(如文本、图像、视频、音频等)。含有丰富表示的多模态数据为网民提供便捷信息的同时，也引发了基于多模态数据的有害信息泛滥。传统的有害信息检测方法主要依赖特征匹配、关键词过滤等技术，在一定程度上可以识别明显的有害信息。然而，除了显性直观的黄色、毒品和暴力外，有害信息还包括潜在的具有隐喻特性的侮辱性、歧视性或令人不安的有害信息。这些隐喻在互联网上随处可见，通过隐含的比喻、类比或象征，将抽象概念或复杂思想转化为更具体、更生动的形象[1]。根据 Lakoff 等[2]的隐喻理论可知，隐喻不仅是语言现象，更是一种思维活动，具有启发理解、说服他人的目的。由于讽刺、隐喻、暗示等多模态隐喻表达方式使得有害信息不直接可见，需要通过多层含义的传递分析和更加复杂的检测才能识别出来，因此，检测多模态隐喻对维护互联网内容安全意义重大。

尽管学者们开始将图像特征和文本特征结合起来，提出多模态隐喻识别方案[3]，但是，现有方案仍面临缺乏数据集和忽略不同模态数据间信息关系不足的问题。针对这一问题，本文从多个角度标注构建数据集，建立注意力机制模型，捕获不同模态数据间的信息关系，以缓解上述不足对隐喻检测研究的影响。该文章的主要内容概述如下：

(1)创建了一个新型具有情感和意图表达的隐喻数据集(metaphor dataset with emotion and intention，MDEI)。收集广告、艺术作品、政治漫画、新闻中的多模态数据，对图像-文本中隐喻出现、情感表达和作者意图进行了详细的标注，以确保 MDEI 的质量。Kappa 分数计算结果表明，所建 MDEI 具有高度的标注一致性。

(2)提出一种基于注意力机制的多模态隐喻检测模型，用于验证 MDEI 中多模态信息间相互关系对模型性能的影响。对比实验结果表明，本文模型的 F1 分数优于基线模型，能为检测任务生成具体的信息量和丰富的语义，数据集对比结果证实了 MDEI 的研究价值。

2　数据集构建

本文构建的 MDEI 旨在支持多模态隐喻检测任务，并为相关研究提供基准和评估平台。数据集包含多种类型的数据，如文本、图像及图像属性信息。文本数据中包含大量的文本语料，

涵盖不同的主题和语境，并包含各种隐喻表达。图像数据中包含与文本相关联的图像信息，用于补充文本信息，提供更丰富的语境。除了原始图像外，数据集还提供与图像相关的实体对象属性信息。这些属性信息涵盖更多的语义信息，而不仅仅是颜色、形状或亮度，用于帮助模型更好地理解图像内容，并与文本进行关联。每个样本均有相应的标签，包括图像名、隐喻出现、源域、目标域、源域模态、目标域模态、隐喻类别、数据来源、情感类别、作者意图、内嵌文本。这些标签可帮助评估模型在隐喻检测任务上的性能。MDEI 将公开发布用于研究。数据集的统计信息如表 1 所示。

表 1 MDEI 的统计信息

Table 1 Statistical information of the MDEI

数据集	训练集	验证集	测试集
隐喻	13 687 对	1 390 对	1 289 对
非隐喻	12 907 对	1 303 对	1 291 对
全部	26 594 对	2 693 对	2 580 对

2.1 数据来源和类型

MDEI 包含 31 867 对图像-文本，其中每对都由一张图像和与之相关的多个文本属性标注信息组成。图像共 1 850 张，覆盖多个类别和场景；文本 1 850 条，包含不同长度和类型的描述。MDEI 可为多模态隐喻领域的研究提供一定的基础数据资源。数据集的文本信息以 JSON 格式存储，便于阅读和编写。图像信息以 PNG 和 IPEG 格式存储。

MDEI 从多种源(如广告、艺术作品、新闻)获取。数据采集主要采用网络爬虫、API 接口调用、数据导入、调查等方法。在采集过程中，重点是收集多样化的模态数据表示(至少具有两种模态)。这些不同模态的数据可提供不同的视角和语境，有助于提高多模态隐喻检测的效果。

(1)广告：广告是一种富有创意的媒介，充满了隐喻语言的运用，是收集多模态隐喻数据的主要来源。

(2)艺术作品：不同艺术家和不同文化背景的艺术作品可为多模态隐喻检测研究提供更全面的视角。

(3)新闻：新闻报道中不仅有文本隐喻，还有图片、视频、音频等多种表现形式的隐喻。不同模态的表现形式可提供更加全面和丰富的隐喻信息。

综上所述，收集多模态隐喻数据集需要广泛的资源和跨学科的合作。除了上述来源外，社交媒体、文学作品等也可作为多模态隐喻数据集的来源。

2.2 数据标注

高质量的数据标注有助于多模态隐喻检测的研究。在参考文献[4]的基础上，本文从图片名、隐喻出现、源域、目标域、源域模态、目标域模态、隐喻类别、数据来源、情感类别、作者意图、内嵌文本等角度开展标注工作。标注示例如图 1 所示。

2.2.1 源域和目标域

在多模态隐喻数据集中，源域和目标域的定义至关重要，因为它们确定了用于检测和理解隐喻的文本和图像元素。源域通常指能够表达抽象概念或隐喻的元素。源域的任务是为模型提供起点，使其能识别出隐喻信息，并理解其中蕴含的意义。目标域通常指用来验证和检测隐喻是否存在的元素。目标域的任务是确保模型能正确理解源域中的隐喻，并将多模态信息(如图像、文本等)相关联，以确认隐喻存在。

2.2.2 隐喻类别

多模态隐喻分为 3 类：文本主导型、图像主导型和互补型。

文本主导型的隐喻主要通过文本信息体现隐喻的出现，图像仅对文本信息进行可视化的补充说明。如图 2(a)所示，文本信息为她头发的颜色像雪花一样，将白发比作雪花。

图片：1.jpg
出现：隐喻
源域：牛奶、奥利奥
目标域：人
目标模态：图像
源域模态：文本、图像
隐喻类别：互补型
数据来源：广告
情感类别：中性
目的：劝说性
内嵌文本：DUNK WITH OREO

文本：吃奥利奥、喝牛奶可以让我离篮球框更近

图 1　图像-文本对的标注示例

Fig. 1　Example of the annotation of the image-text pair

(a) 她头发的颜色像雪花一样　(b) 城市逐渐变成了灰色　(c) 燃烧的橙子，热情似火

图 2　多模态隐喻类别样例

Fig. 2　Samples of the multi-modal metaphor category

图像主导型的隐喻指图像本身实现了隐喻信息的传达。如图 2(b)所示，吸烟使得原本色彩斑斓的城市变成灰烬。而文本信息“城市逐渐变成了灰色”则进一步加强了这种关系，使隐喻更加容易检测。

互补型的隐喻通过文本和图像的共同内容体现隐喻的信息，也就是“图像-文本对”之间的相互关系。在互补型隐喻中，单凭文字或者图片无法确定“图像-文本对”是否包含隐喻信息。图 2(c)是一个点燃的橙子，并配以文本“燃烧的橙子，热情似火”，将橙子的形象与充满活力的火焰相联系。鲜艳的橙色与火焰的熊熊燃烧相互映衬，传递出“激情、活力、能量”等内涵，突显了其具有充沛活力和无限激情之特质。

数据集的隐喻类别创建通过人工标注完成，前期通过模型对数据集隐喻类别进行训练，后期主要作为分类识别的任务。后期识别利用机器学习技术构建模型实现自动识别。后期的识别方式能充分发挥人工标注的优势和学习的自动化能力，提高隐喻类别的效率和准确性。

2.2.3　情感类别和作者意图

通过巧妙设计，互联网上的多模态隐喻能更加含蓄地表达情感，并传递作者意图，潜移默化地影响网民，并使其被说服。因此，在多模态隐喻信息中，理解情感类别和作者意图具有重要意义。理解情感类别和作者意图需要全面分析和解读隐喻中的多模态信息(如语言、视觉元素、音频等)。情感类别通常指文本或图像传达的情感色彩(如积极、中立、消极)，而作者意图则是作者想要通过多模态传达的信息或目的。作者意图主要分为表达性、描述性和说服性 3 类。表达性指多种模态的数据结合能表达作者的情感、态度、思想、感觉。描述性指对实体、事件、概念、信息、动作和角色的描述，用以丰富或扩大

隐喻语义。说服性指在广告和社交媒体等沟通环境中，利用隐喻说服读者购买某物品或采取某些行动。

2.2.4 外部知识

通常，隐喻背后蕴含特定的文化背景和常识，因此需要通过引入外部知识提升数据集质量，为模型提供更丰富的图像描述信息，从而弥合多模态数据间存在的模态差距。本文利用基于预训练的 ClipCap 模型[5]的自动化实现方法生成具有丰富语义信息的描述性标题，并将其应用于每张图像上。通过引入图像标题等外部知识，可为 MDEI 增加更多上下文语义，并扩充多模态数据集。

2.3 数据一致性评估

多模态隐喻的标注通常和标注者的主观判断和先验知识有关，因此可能导致数据存在标注者的主观偏见特性。在多模态隐喻数据的标注中，不同标注者可能会因为主观判断和先验知识不同而产生不同的标注结果，因此需要量化他们之间的一致性水平。

参考文献[6]中使用的评估方式，本文同样利用 Kappa 分数[7]评估标注者间的一致性程度。Kappa 分数(k)是一种用于度量标注者间一致性的统计方法，特别适用于标注者在无法确定正确标注时选择一个随机性标注的情况。Kappa 分数的计算基于混淆矩阵(该混淆矩阵是一个表格)，用于比较两个标注者的标注结果是否一致。混淆矩阵通常包括 4 个值：*TP* 为标注者们都正确标注的样本数量；*TN* 为标注者们都未正确标注的样本数量；*FP* 为一个标注者标注了某个样本属于某一特定类别(或标注了某个特定属性)，但另一个标注者认为该样本不属于该类别(或未标注该属性)的样本数量；*FN* 为一个标注者未标注某个样本属于某一特定类别(或未标注某个特定属性)，但另一个标注者认为该样本属于该类别(或标注了该属性)的样本数量。计算 Kappa 分数的步骤如下。

首先，依据式(1)计算实际一致性 p，用于定义标注者们实际观察到的一致性。

$$p=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

其次，依据式(2)计算期望一致性 p_e。式(2)假设标注完全随机的情况下，评估标注者们之间的一致性。

$$p_e=\frac{(TP+FP)\times(TP+FN)+(FN+TN)\times(FP+TN)}{(TP+TN+FP+FN)} \tag{2}$$

最后，利用式(3)计算 Kappa 分数。

$$k=\frac{p-p_e}{1-p_e} \tag{3}$$

Kappa 分数计算完成后，利用表 2 所示的评估标准衡量评估标注者间一致性的程度。

表 2 Kappa 分数的评估标准

Table 2 Evaluation criteria for the Kappa scores

Kappa 分数	标注一致性
0.00～0.20	极低的一致性
0.21～0.40	一般的一致性
0.41～0.60	中等的一致性
0.61～0.80	高度的一致性
0.81～1.00	几乎完全一致

依据式(1)～(3)，对数据集 MDEI 进行计算，发现隐喻识别一致性得分为 0.77，表明标注者对隐喻识别具有较高的一致性。采用类似方式，对数据集的隐喻类型识别、源域和目标域所属模态识别、情感类别及作者意图类别进行一致性得分，其结果分别为 0.74、0.75、0.83、0.82。结果表明，本文所建数据集 MDEI 具有高度的标注一致性。

2.4 数据管理和维护

数据集的规模和质量影响多模态隐喻检测效果。数据集规模越大，覆盖的模态表达形式和语境越多，越可提高多模态隐喻检测的准确性和鲁棒性。为便于数据管理和维护，实现高质量的 MDEI 标注，本文做了如下工作：定义数据收集

标准，以确保数据的代表性和多样性；采用多人标注、验证等方式，以保证数据的真实性和准确性；做好数据内容的安全隐私保护。

由于多模态隐喻现象的多样性和复杂性，且随着时间的推移，新的隐喻表达方式和语境将不断涌出，因此需要定期更新和维护数据集，以保证其具有实时性和有效性。

3　基于注意力机制的数据集验证方法

鉴于注意力机制可关注多模态数据的重要特征，因此本节建立一个基于注意力机制的多模态隐喻检测模型(图 3)，以验证不同模态间的交互关系对模型的影响，进一步评估所建数据集的研究价值和质量。模型借助残差神经网络(Residual neural Network，ResNet)、双向长短期记忆(Bi-Long Short Term Memory，Bi-LSTM)、YOLOV7+ 全局词向量表示(Global Vectors for Word Representation，Glove)等技术手段，从不同的模态中提取语义特征。首先，模型将图像特征向量和文本特征向量映射到相同的维度空间。其次，利用注意力机制计算图像和文本特征之间的权重。再次，利用注意力权重将图像特征与文本特征向量进行加权融合。最后，利用 *softmax* 全连接层对融合特征进行分类，得到多模态隐喻预测结果，即判断“图像-文本对”是否含有隐喻。

3.1　特征提取

模型主要提取文本特征、图像属性特征、图像实体对象特征。文本特征主要捕获数据标注信息和基于预训练模型生成的图像标题，然后再借助预训练模型[8]和 Bi-LSTM 网络[9]进行特征学习。在图像特征提取上，借助 ResNet-50 V2[10]图像编码器捕获图像的颜色、形状、纹理和空间位置关系等特征。图像通常包括多个实体对象，实体对象对隐喻语义的理解有利，主要通过以下方式获得图像实体对象特征。首先，图像实体对象基于 YOLOV7 和 ImageNet 图像数据集训练一个实体预测器；其次，利用预测器生成概率最高的 5 个实体对象，并用其实体名称作为实体对象标签；最后，借助预训练模型将标签转换为图像实体对象向量，进而实现图像实体对象特征的提取。

3.2　特征融合

多项研究证实，注意力机制能自动学习数据特征之间的关联，并根据重要性加权融合不同特征，以丰富数据表示。因此，本文构建一个基

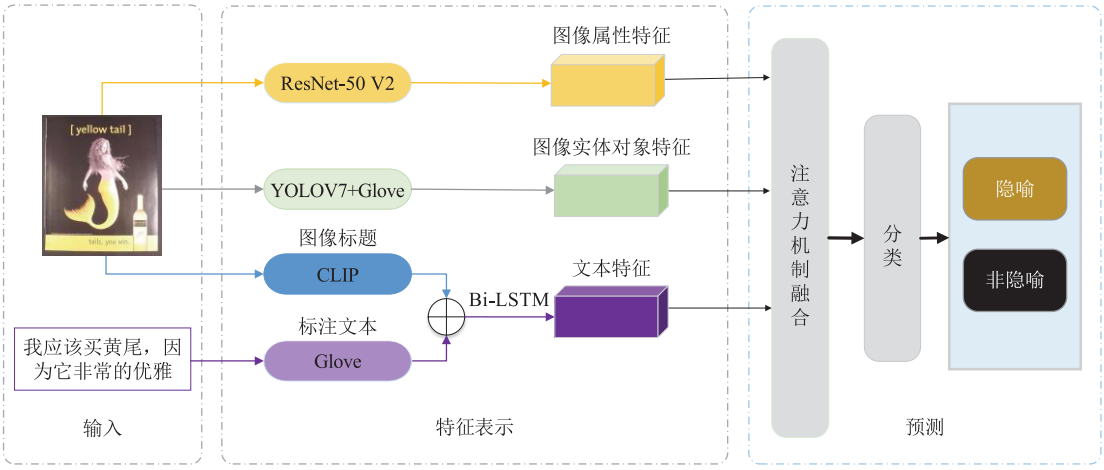

图 3　基于注意力机制的评估模型

Fig. 3　Evaluation model based on attention mechanism

于注意力机制的隐喻检测模型，捕获不同模态数据间信息关系，以提高隐喻检测的准确性和鲁棒性。模型利用每个模态的特征，并根据模态的重要性进行加权处理，从而得到全面综合的隐喻检测特征表示。首先，将每个模态 n 的特征向量 $\boldsymbol{v}_n$ 转换为定长形式 $\boldsymbol{v}_n'$。其次，通过两层前馈神经网络计算每个模态的注意力权重，并将这些权重应用于特征向量 $\boldsymbol{v}_n'$，加权平均计算得到固定长度的向量 $\boldsymbol{v}_{fused}$，如式(4)～(6)所示。最后利用 *softmax* 激活的全连接层进行分类，得到多模态隐喻预测结果。

$$\boldsymbol{\omega}_n'=\boldsymbol{W}_{n2}\cdot\tanh\left(\boldsymbol{W}_{n1}\cdot\boldsymbol{f}_n+\boldsymbol{b}_{n1}\right)+\boldsymbol{b}_{n2} \tag{4}$$

$$\boldsymbol{v}_n'=\tanh\left(\boldsymbol{W}_{n3}\cdot\boldsymbol{v}_n+\boldsymbol{b}_{n3}\right) \tag{5}$$

$$\boldsymbol{v}_{\text{fused}}=\sum\nolimits_{n\in\{\text{text, image, attr}\}}\boldsymbol{\omega}_n'\boldsymbol{v}_n' \tag{6}$$

其中，n 为 3 种模态之一；$\boldsymbol{\omega}$ 为包含 $\boldsymbol{\omega}_n'$ 的矢量；$\boldsymbol{W}_{n1}$、$\boldsymbol{W}_{n2}$、$\boldsymbol{W}_{n3}$ 为对应模态的权重矩阵；$\boldsymbol{b}_{n1}$、$\boldsymbol{b}_{n2}$、$\boldsymbol{b}_{n3}$ 为偏差；$\boldsymbol{f}_n$ 为融合过程重构的特征向量。

3.3 实验设置

本文构建的数据集 MDEI 的多模态数据大于 1 300 对，每对均由一张图像和相应标注信息组成。采用 PyTorch[11]开发模型，利用 Glove 技术[12]在数据集上训练生成词嵌入和属性嵌入。在模型训练中，将数据集按 8∶1∶1 比例划分为训练集、验证集和测试集，并设置词向量维度为 512，属性嵌入维度为 200。同时采用自适应矩阵估计法(Adam)优化器[13]及梯度裁剪方法[14]进行训练，以避免梯度爆炸问题。其他超参数设置如表 3 所示。

表 3 超参数

Table 3 Hyperparameter

Hyper-parameters	Value
LSTM hidden size	256
Batch size	64
Learning rate	0.000 1
Gradient Clipping	5
Early stop patience	5
ResNet FC size	512
LSTM dropout rate	0.2
Classification layer 12 parameters	1×10^{-7}

3.4 结果分析

模型除了对比全部非隐喻、随机预测外，还包括仇恨检测任务[15]、讽刺检测任务[16]和文本隐喻检测任务[17]。对比指标采用融合精确率和召回率的 F1 分数，F1 分数越高，性能越好。本文模型和其他模型间的对比结果如表 4 所示。

由表 4 可知，本文模型在多模态检测任务中表现最好，F1 分数达到了 93.68%。文本隐喻检测数据集虽然相对丰富，且模型较为健全，但缺乏图像数据信息，无法对隐喻的源域和目标域进行模态定位。由于多模态隐喻检测任务比其他多模态分类任务更复杂，所以讽刺识别和仇恨检测中的两个模型在隐喻检测任务中的表现也不如本文模型。本文数据集虽然参考 Zhang 等[4]的收集

表 4 模型对比结果

Table 4 Model comparison results

模型	准确率（%）	精确率（%）	召回率（%）	F1 分数
全部非隐喻	50.42	-	-	-
随机预测	52.19	53.89	49.32	51.50
Gomez 等[15]	85.67	84.02	78.15	80.98
Liu 等[16]	86.28	82.42	85.67	84.01
Lin 等[17]	81.32	78.82	79.73	79.27
Multi-MET[5]	87.19	89.77	90.57	90.17
本文	87.32	94.91	92.49	93.68

思路，但是本文模型性能优于 Zhang 等[4]提出的模型。因为 Zhang 等[4]提出的模型仅将图像和文本特征结合在一起，而本文模型引入图像实体对象，作为连接图像和文本的附加模态。

本文对 MDEI 和其他多模态隐喻数据集进行对比，评价指标为 F1 分数，结果如表 5 所示。由表 5 可知，MDEI 的性能较优，F1 分数达到了 93.57%，表明 MDEI 可提升多模态隐喻检测任务的性能。

上述实验表明，本文构建的数据集 MDEI 能提升模型检测效果，所提出的注意力机制模型能捕获不同模态之间的关联关系，进而能更好地理解和感知多模态数据中的隐喻含义。数据集 MDEI 的构建和验证工作对开展多模态隐喻研究有利。

4　研究现状

近年来，文本隐喻检测领域涌现了多个数据集。VUA[20]是最早出现的文本隐喻数据集之一。Tsvetkov 等[21]基于网络资源构建了一个包含 2 000 个“形容词-名词”对的数据集。Mohammad 等[22]在语言层面进行了词级别标注，他们提出的数据集共包含 761 个句子。Steen 等[23]提出隐喻检测指南，以规范隐喻标注过程。Shutova 等[24]创建了一个包含情感和政治色彩的隐喻数据集。Zayed 等[25]实施了数据集的词级别隐喻标注。

部分研究者开始关注多模态隐喻数据集的构建。Shutova 等[18]通过利用给定短语进行谷歌搜索获得“图像-文本”组合方式。Steen[19]构建了一个在线、可动态扩充的多模态隐喻数据集 VisMet。Zhang 等[4]提出一个可定量研究多模态相互作用下隐喻检测问题的数据集。

一些研究者开始探索多模态隐喻检测方法。Kehat 等[26]提出基于视觉词嵌入的多模态隐喻检测方法。Zhang 等[4]提出情感检测和作者意图检测模型。Li 等[27]通过文本描述和大规模情绪语言实现预测。Zhang 等[28]提出一种自适应特定模态权值融合网络，以解决多模态数据融合过程中的问题。

5　结论和展望

本文从多个角度标注并构建了一个多模态隐喻数据集，评估了一致性标注程度，建立了一个基于注意力机制的多模态隐喻检测模型，以研究多模态信息间关系对隐喻的影响，结果表明，模型和数据集的指标均优于对比基线。多模态隐喻数据集虽然能为研究者开展隐喻研究工作提供数据支撑，但高质量大规模多模态隐喻数据集的创建需要大量专业知识与标记工作。因此，未来工作计划从数据标注、数据合成和主动学习等方面展开，具体如下：扩充数据标注内容，进一步标记物体位置、文本情感和歧义信息；利用对抗生成网络生成虚拟多样化情境和隐喻类型样本；采用主动学习策略自动标注不确定性样本，以优化数据集质量。同时，探索音频、视频等其他模态对多模态隐喻的影响。

表 5　数据集对比结果

Table 5　Comparison results of the dataset

数据集	准确率（%）	精确率（%）	召回率（%）	F1 分数
Shutova 2016[18]	82.56	80.36	77.25	78.77
VisMet[19]	83.17	81.53	77.98	79.72
Multi-MET[4]	87.19	89.77	90.57	90.17
MDEI	87.39	94.98	92.55	93.75

参 考 文 献

[1] Aggarwal S, Singh R. Metaphor detection using deep contextualized word embeddings [Z/OL]. arXiv Preprint, arXiv: 2009.12565, 2020.

[2] Lakoff G, Johnson M. Metaphors we live by [M]. Chicago: University of Chicago Press, 1980.

[3] Shutova E, Sun L, Korhonen A. Metaphor identification using verb and noun clustering [C] // Proceedings of the 23rd International Conference on Computational Linguistics, 2010: 1002-1010.

[4] Zhang DY, Zhang MH, Zhang HT, et al. Multimet: a multimodal dataset for metaphor understanding [C] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 3214-3225.

[5] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.

[6] 张明昊. 基于双线性池化的多模态隐喻识别 [D]. 大连: 大连理工大学, 2022.
Zhang MH. Multimodal metaphor identification based on bilinear pooling [D]. Dalian: Dalian University of Technology, 2022.

[7] Fleiss JL. Measuring nominal scale agreement among many raters [J]. Psychological Bulletin, 1971, 76(5): 378-382.

[8] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [Z/OL]. arXiv Preprint, arXiv: 1810.04805, 2018.

[9] Medsker LR, Jain LC. Recurrent neural networks [J]. Design and Applications, 2001, 5: 64-67.

[10] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[11] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library [Z/OL]. arXiv Preprint, arXiv: 1912.01703, 2019.

[12] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation [C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.

[13] Kingma DP, Ba LJ. Adam: a method for stochastic optimization [C] // Proceedings of the 2015 International Conference on Learning Representations, 2015: 13.

[14] Zhang JZ, He TX, Sra S, et al. Why gradient clipping accelerates training: a theoretical justification for adaptivity [C] // Proceedings of the 8th International Conference on Learning Representations, 2020: 1-21.

[15] Gomez R, Gibert J, Gomez L, et al. Exploring hate speech detection in multimodal publications [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 1470-1478.

[16] Liu H, Wang WY, Li HL. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement [Z/OL]. arXiv Preprint, arXiv: 2210.03501, 2022.

[17] Lin ZX, Ma QL, Yan JY, et al. CATE: a contrastive pre-trained model for metaphor detection with semi-supervised learning [C] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 3888-3898.

[18] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: metaphor identification with visual features [C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 160-170.

[19] Steen GJ. Visual metaphor: structure and process [M]. Netherlands: John Benjamins Publishing Company, 2018.

[20] Birke J, Sarkar A. A clustering approach for the nearly unsupervised recognition of nonliteral language [C] // Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006: 329-336.

[21] Tsvetkov Y, Boytsov L, Gershman A, et al. Metaphor detection with cross-lingual model transfer [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 248-258.

[22] Mohammad SM, Shutova E, Turney PD. Metaphor as a medium for emotion: an empirical study [C] // Proceedings of the 5th Joint Conference on Lexical and Computational Semantics, 2016: 23-33.

[23] Steen GJ, Dorst AG, Herrmann JB, et al. A method for linguistic metaphor identification: from MIP to MIPVU [M]. Netherlands: John Benjamins Publishing Company, 2020.

[24] Shutova E, Teufel S. Metaphor corpus annotated for source-target domain mappings [C] // Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010: 3255-3261.

[25] Zayed O, McCrae JP, Buitelaar P. Crowd-sourcing a high-quality dataset for metaphor identification in tweets [C] // Proceedings of the 2nd Conference on Language, Data and Knowledge, 2019, 10:1-10:7.

[26] Kehat G, Pustejovsky J. Integrating vision and language datasets to measure word concreteness [C] // Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017: 103-108.

[27] Li SX, Okada S. Interpretable multimodal sentiment analysis based on textual modality descriptions by using large-scale language models [Z/OL]. arXiv Preprint, arXiv: 2305.06162, 2023.

[28] Zhang JL, Wu XM, Huang CQ. AdaMoW: multimodal sentiment analysis based on adaptive modality-specific weight fusion network [J]. IEEE Access, 2023, 11: 48410-48420.