一种多模态隐喻数据集的构建和验证方法

夏冰[1,2]，杨瑞楠[1,2]，董玉[1,2]，楚世豪[1,2]，唐崇俊[1,2]，葛云翔[1,2]，尹家斌[1,2]

[1]（中原工学院前沿信息技术研究院，郑州，450007）

[2]（河南省网络舆情监测与智能分析重点实验室，郑州，450007）

摘要：隐喻具有启发理解、说服他人的目的。当前隐喻呈现文本、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，当前学者们难以建立研究模型，使其更多关注基于文本的隐喻检测。针对这一不足，本文首先从图像-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（Metaphor Dataset with Emotion and Intention, MDEI）。接着，利用 Kappa 分数评估数据集标注者间的一致性。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，MDEI 能提升隐喻模型检测效果，多模态信息间相互关系能有助于隐喻的理解。

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

Bing Xia[1,2],Ruinan Yang[1,2],Yu Dong[1,2],Shihao Chu[1,2],Chongjun Tang[1,2],Yunxiang Ge[1,2],Jiabin Yin[1,2]

[1] (Zhongyuan University of Technology, Zhengzhou, 450007, China)

[2](Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou, 450007, China)

Abstract: Metaphor has the purpose of inspiring understanding and persuading others. At present, metaphor presents the trend of multimodal integration of text, image, and video. Therefore, identifying the metaphorical semantics contained in multimodal contents is of research value for Internet content security.. Due to the lack of multimodal metaphor data sets, it is difficult for scholars to build research models and pay more attention to text-based metaphor detection. To overcome this shortcoming, we first generate a new multimodal metaphor dataset MDEI from the perspectives of image-text, metaphor appearance, emotion expression, and author intention. Then, Kappa scores were used to assess the consistency among the 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 MDEI can improve the effectiveness of metaphor model detection, and confirm that the interrelationship of multimodal information is helpful for understanding metaphor.

Key words: Content security; multimodal metaphor detection; external knowledge; multi-modal dataset; attention mechanism.

来稿日期：2024-01-24 修回日期：2024-04-28

基金项目：河南省科技攻关项目（232102211088）和河南省哲学社会科学规划项目（2022BXW018）

作者简介：夏冰，博士，副教授，CCF 高级会员，研究方向为网络安全和网络舆情，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，网络安全。楚世豪，硕士研究生，网络安全。唐崇俊，硕士研究生，网络安全。葛云翔，硕士研究生，网络安全。尹家斌，硕士研究生，网络安全。

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），含有丰富表示的多模态为网民提供便捷信息的同时，也引发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于特征匹配、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，有害信息除了显性直观的黄色、毒品和暴力之外，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 和 Johnson 的隐喻理论[2]，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]，然而，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，我们从多个角度标注来构建数据集，建立注意力机制模型捕获不同模态数据间信息关系，以缓解上述不足对隐喻检测研究的影响。本文主要贡献如下：

(1) 创建了一个新型多模态隐喻数据集 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角度，进行了详细的标注工作以确保 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，模型 F1 分数优于基线模型，能够为检测任务生成具体的信息量和丰富的语义，数据集对比结果也证实了 MDEI 的研究价值。

2 数据集构建

本文构建的 MDEI 数据集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。数据集包含了多种类型的数据，包括文本、图像以及图像属性信息。文本数据中包含了大量的文本语料，涵盖不同的主题和语境，并包含各种隐喻表达。图像数据中包含了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。除了原始的图像外，数据集还提供了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。每个样本都会有相应的标签，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。这些标签可以帮助评估模型在隐喻检测任务上的性能。MDEI 数据集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 数据集的统计信息

Table1 Statistical Information of the MDEI Dataset

	Training	Valid	Test
Metaphor	13687	1390	1289
Non-metaphor	12907	1303	1291
All	26594	2693	2580

2.1 数据来源和类型

MDEI 数据集包含 1850 个图像-文本对，其中每个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。图像数量共计 1850 张,覆盖了多个类别和场景;文本数量同样为 1850 条，包含了不同长度和类型的描述。MDEI 数据集可以为多模态隐喻领域的研究提供一定的基础数据资源。数据集的文本信息是以 JSON 格式存储，便于阅读和编写，也易于机器解和生成。图像信息是以 PNG 和 IPEG 的格式存储。

MDEI 从多种源(如:广告、艺术作品、新闻)获取。数据采集土要是采用网络爬虫，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（如：至少具有 2 种模态）。这些不同模态的数据可以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，可以为多模态隐喻检测研究提供更全面的视角。

（3）新闻：在新闻报道中，不仅存在着文本隐喻，还存在着图片、视频、音频等多种表现形式，这些多模态元素可以提供更加全面和丰富的隐喻信息。

总之，收集多模态隐喻数据集需要广泛的资源和跨学科的合作。除了上述来源之外，社交媒体、文学作品等也可以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[5]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，内嵌文本角度开展标注工作。图 1 是一个标注示例。下面介绍隐喻类别、情感类别、作者意图等重要标注。

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：互补性
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

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，使其能够识别出隐喻信息并理解其中蕴含的含义。目标域通常指用来验证和检测隐喻是否存在所需的元素。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

多模态隐喻分为三类：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来体现隐喻的出现，图像只是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“她头发的颜色像雪花一样”，将白发比作雪花。

图像主导型的隐喻是指图像本身实现了隐喻信息的传达。如图 2（b）所示，从图中可以看到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。

互补型的隐喻是通过文本和图像共同内容来体现隐喻的信息，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。如图 2（c）所示，展现了一个点燃的橙子，并配以文本“燃烧的橙子，热情似火”，将橙子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形成强烈对比，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。

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

数据集的隐喻类别创建是通过人工标注完成的，后期是作为分类任务进行识别。后期识别打算利用机器学习技术构建模型实现自动识别。这种方式能够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够更加含蓄地表达情感并传递作者意图，在潜移默化之间影响着网民并使其被说服。因此，在多模态隐喻中，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）。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、描述性和说服性三类。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。描述性实现对实体、事件、概念、信息、动作和角色的描述，以丰富或扩大隐喻语义。说服性则是在广告和社交媒体等沟通环境中，使用隐喻来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，隐喻背后蕴含着特定的文化背景和常识，因此需要引入外部知识来提升数据集质量，为模型提供更加丰富的图像描述信息，从而弥合多模态数据之间存在的模态差距。本文采用一种基于预训练的 Clipcap (CLIP Prefix for Image Captioning, Clipcap)模型[4]的自动化实

现方法，利用该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，可以为 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。

类似论文[6]，我们使用 Kappa 分数[7]来评估标注者间的一致性程度。Kappa（k）分数是一种用于度量标注者之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。混淆矩阵通常包括四个值，True Positives (TP)表示标注者们都正确标注的样本数量，True Negatives (TN)表示标注者们都正确未标注的样本数量，False Positives (FP)表示一个标注者标注了但另一个标注者没有标注的样本数量，False Negatives (FN)表示一个标注者没有标注但另一个标注者标注了的样本数量。具体计算 Kappa 分数的步骤如下：

首先，依据公式 (1) 计算实际一致性 (p)，该公式用于定义标注者们实际观察到的一致性。

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

然后，依据公式 (2) 计算期望一致性 (p_e)，该公式假设标注是完全随机的情况下，评估标注者们之间的一致性。

$$P_e = \frac{[(TP + FP) * (TP + FN) + (FN + TN) * (FP + TN)]}{[(TP + TN + FP + FN)]} \tag{2}$$

最后，使用 Kappa 分数的计算公式 (3) ，计算 Kappa 分数。

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

Kappa 分数计算完成后，利用表 2 所示的学术界主流的评估标准来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	标注一致性水平
<0	完全不一致
0.01-0.20	极少数不一致
0.21-0.40	少数一致
0.41-0.60	部分一致
0.61-0.80	大部分一致
0.81-1.00	基本完全一致

依据上述公式，本文在数据集 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，MDEI 也计算了隐喻类型识别、源域和目标域所属模态识别、情感类别以及作者意图类别的一致性得分，其结果分别为 0.74、0.75、0.83 和 0.82。多个结果表明，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，表明覆盖的模态表达形式和语境越多，进而能提高多模态隐喻检测的准确性和鲁棒性。数据集中的样本数据应该具有代表性和多样性。为了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

鉴于注意力机制能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（如图 3 所示），以验证不同模态之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。模型首先将图像特征向量和文本特征向量映射到相同的维度空间。接着，使用注意力机制来计算图像和文本特征之间的权重。然后，使用其注意力权重，将图像特征与文本特征向量进行加权融合。最后，使用 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。

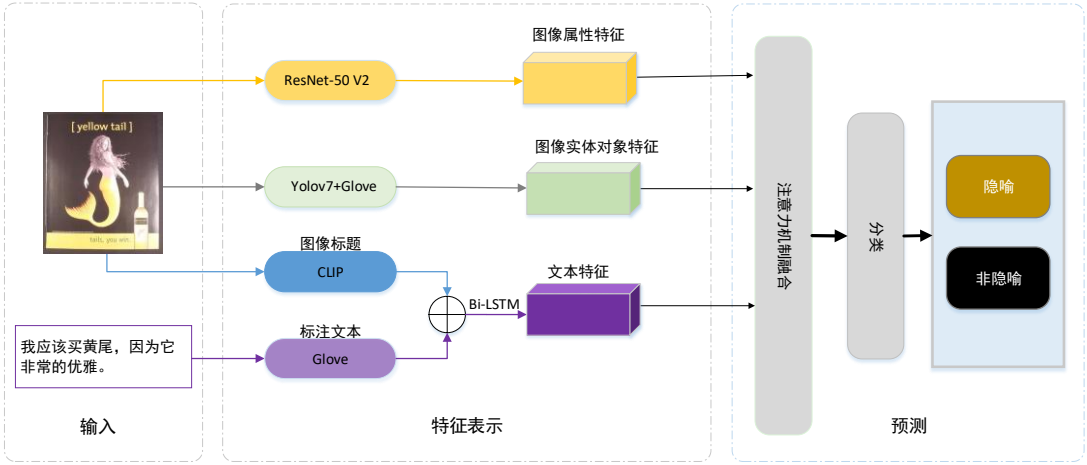

图 3 基于注意力机制的评估模型

Fig.3. Evaluation model based on attention mechanism

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-50v2[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，这些对象对隐喻语义的理解具有帮助，因此为了获得图像实体对象特征。图像实体对象首先基于 YOLOV7 和 ImageNet 图像数据集训练一个实体预测器，接着利用预测器生成概率最高的五个实体对象，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，从而得到全面的综合的隐喻检测特征表示。首先，将每个模态n的特征向量v_n转换为定长形式v_n'。然后，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v_n'，加权平均计算得到固定长度的向量v_{fused}（如公式（4-7）所示）。最后使用 softmax 激活的全连接层进行分类，得到多模态隐喻预测结果。

$$\omega_n' = W_{n2} \cdot tanh(W_{n1} \cdot f_n + b_{n1}) + b_{n2} \tag{4}$$

$$\omega = softmax(\omega) \tag{5}$$

$$v_n' = tanh(W_{n3} \cdot v_n + b_{n3}) \tag{6}$$

$$v_{fused} = \sum_{n \in \{text, image, attr\}} \omega_n' v_n' \tag{7}$$

其中，n 是三种模态之一，而ω是包含ω_n' 的矢量；W_{n1} W_{n2} W_{n3}是对应模态的权重矩阵。b_{n1} b_{n2} b_n是偏差; f_n是表示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 超过 1300 对多模态数据，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(Adaptive Moment Estimation, 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	0.0001
Gradient Clipping	5
Early stop patience	5
ResNet FC size	512
LSTM dropout rate	0.2
Classification layer 12 parameters	1e -7

3.4 结果分析

模型除了对比全部非隐喻、随机预测之外，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，表示性能越好。本文模型和其他模型之间的对比结果如表 4 所示。

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
全部非隐喻	50.42	-	-	-
随机预测	52.19	53.89	49.32	51.50
Gomez 等人[15]	85.67	84.02	78.15	80.98
Liu 等人[16]	86.28	82.42	85.67	84.01
Lin 等人[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**

从表 4 中的对比结果可以看出，本文模型在多模态检测任务表现最好，F1 分数达到了 93.68。尽管文本隐喻检测数据集相对丰富且模型较为健全，但由于缺乏图像数据信息，无法对隐喻的源域和目标域进行模态定位。面向讽刺识别和仇恨检测多模态任务中两个模型在隐喻检测任务中表现也不如本文模型，因为多模态隐喻检测任务比其他多模态分类任务更复杂。尽管本文数据集参考了 Zhang 等人[5]的收集思路，但是本文模型性能各项指标均要高于其提出的模型。因为其模型只是将图像和文本特征结合在一起，而本文模型引入了图像实体对象作为连接图像和文本的附加模态。

本文还实现了 MDEI 和其他学者提出的多模态隐喻数据集之间的对比，对比评价指标仍然采用 F1 分数，对比结果如表 5 所示。从数据集对比结果来看，MDEI 在实验中表现出卓越的性能，超过了 Shutova 等人、Steen 等人和 Zhang 等人创建的数据集，F1 分数达到了 93.75，证实了 MDEI 能提升多模态隐喻检测任务的性能。

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
Shutova 2016[25]	82.56	80.36	77.25	78.77
VisuMet[26]	83.17	81.53	77.98	79.72
Multi-MET[5]	87.19	89.77	90.57	90.17
本文	**87.39**	**94.98**	**92.55**	**93.75**

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够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 构建和验证工作，对开展多模态隐喻研究具有价值。

4 研究现状

近年来，在文本隐喻检测领域涌现了多个数据集。TroFi[18]和 VUA[19]是最早出现的文本隐喻数据集之一。Tsvetkov 等人[20]基于网络资源构建了一个包含 2000 个“形容词-名词”对的数据集。Mohamad 等人[21]在语言层面进行了词级别标注，该数据集共包含 761 个句子。

Steen[22]提出了隐喻检测指南，以规范隐喻标注过程。Shutova[23]和 Teufel 创建了一个包含情感和政治色彩的隐喻数据集。Zayed 等人[24]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 等人[25]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 构建了一个在线、可动态扩充的多模态隐喻数据集 VisMet[26]。Zhang 等人[5]提出一个能够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 和 Pustejovsky[27]提出了基于视觉词嵌入的多模态隐喻检测方法。Zhang 等人[5]提出了情感检测和作者意图检测模型。Li 等人[28]通过文本描述和大规模情绪语言实现预测。Zhang 等人[29]提出了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。多模态隐喻是一个相对新的领域，使用多模态隐喻关键词搜索了相关数据库后仅得到了以上相关文献。

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。尽管多模态隐喻数据集能为研究者开展隐喻研究工作提供数据支撑，然而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。在下一步工作中计划从数据标注、数据合成和主动学习等方面展开。具体包括扩充数据标注内容，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。

参考文献

[1] Aggarwal S, Singh R. Metaphor Detection using Deep Contextualized Word Embeddings, 2020. DOI: 10.48550/arXiv.2009.12565.

[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] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of International Conference on Machine Learning Research, 2021: 8748-8763.

[5] 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.

[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.

[7] Zhang MH. Multimodal metaphor identification based on bilinear pooling [D]. Dalian: Dalian University of Technology, 2022.

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

[9] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv:1810.04805, 2018.

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

[11] 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.
[12] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library[J]. arXiv preprint, arXiv:1912.01703, 2019.
[13] 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.
[14] Kingma DP, Ba J. Adam: a method for stochastic optimization [C] // Proceedings of the 2015 International Conference on Learning Representations: 13.
[15] 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: 26-30.
[16] 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.
[17] Liu H, Wang WY, Li HL. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement [J]. arXiv preprint, arXiv:2210.03501, 2022.
[18] Lin Z, Ma Q, Yan J, 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.
[19] Steen G, Dorst AG, Herrmann JB, et al. A method for linguistic metaphor identification [J]. Amsterdam: Benjamins, 2010.
[20] Birke J, Sarkar A. A clustering approach for 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, 1: 248-258.
[22] Mohammad S, Shutova E, Turney P. 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 [J]. Metaphor and the Social World, 2014, 4(1):138-146.
[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.
[26] 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.
[27] Steen GJ. Visual metaphor: structure and process [M]. The Netherlands: John Benjamins Publishing Company, 2018.

[28] 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.
[29] Li S, Okada S. Interpretable multimodal sentiment analysis based on textual modality descriptions by using large-scale language models [J]. arXiv preprint, arXiv:2305.06162, 2023.
[30] Zhang J, Wu X, Huang C. AdaMoW: multimodal sentiment analysis based on adaptive modality-specific weight fusion network [J]. IEEE Access, 2023, 11: 48410-48420.