探索多模式隐喻检测的思维链

Yanzhi Xu,* **Yueying Hua,*** **Shichen Li** and **Zhongqing Wang†**
苏州大学自然语言处理实验室，中国苏州
{yzxuxyz,yyhua1224}@stu.suda.edu.cn scli_21@outlook.com,
wangzq@suda.edu.cn

摘要

隐喻常见于广告和网络流行语中。然而，网络流行语的自由形式往往导致缺乏高质量的文本数据。隐喻识别需要对文本和视觉元素进行深入解读，需要大量常识性知识，这对语言模型提出了挑战。为了应对这些挑战，我们提出了一个名为 C4MMD 的组合框架，它利用思维链（CoT）方法**进行**多模态隐喻检测。具体来说，我们的方法设计了一个受 CoT 启发的三步流程，从多模态大语言模型（MLLM）中提取知识并将其整合到小语言模型中。我们还开发了一种模态融合架构，将大型模型中的知识转化为隐喻特征，并辅以辅助任务来提高模型性能。在 MET-MEME 数据集上的实验结果表明，我们的方法不仅有效地增强了小型模型的隐喻检测能力，而且优于现有模型。据我们所知，这是第一项在隐喻检测任务中利用 MLLM 的系统性研究。我们的方法代码可在 https://github.com/xyz189411yt/C4MMD.

*平等贡献
†通讯作者

1 导言

隐喻在我们的日常表达和写作中非常普遍，会对自然语言处理（NLP）的下游任务产生一系列影响，如语义理解（Neuman 等人，2013 年）、情感分析（Ghosh 和 Veale，2016 年；Mohammad 等人，2016 年）和其他任务。近年来，社交媒体的兴起引发了人们对多模态隐喻的兴趣。因此，出现了多个多模态隐喻数据集。

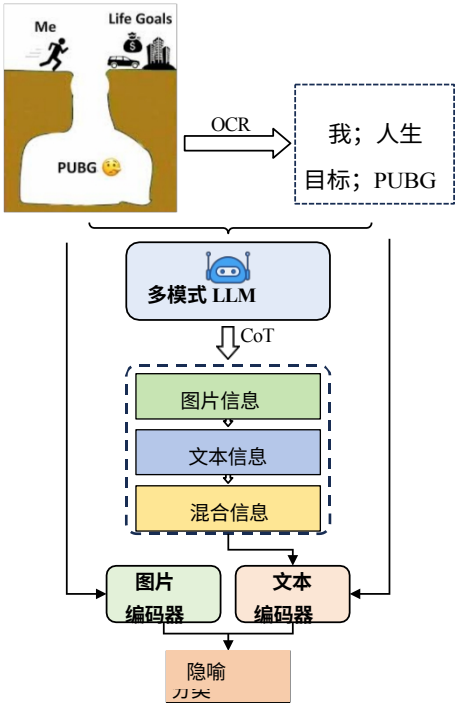


图 1：多模式隐喻检测示例。

有人提出了模态隐喻（Zhang 等人，2021，2023a；Alnajjar 等人，2022）。

目前关于多模态隐喻保护的研究仍处于早期阶段。主要的挑战在于多模态隐喻的复杂性和多样性。与单模态检测相比，多模态隐喻检测不仅能发现句子中的隐喻，还能将其分为图像主导型、文本主导型或互补型。第二个主要挑战来自于文本内容的低质量，这些内容主要来自于社交媒体上的广告和备忘录。文字赋予图像更多的隐喻性。最近的研究使用 OCR（光学字符识别）来提取图像中的文本。然而，仅仅依靠 OCR 将其转换为并行文本会导致文本位置的丢失。

第 62 届计算语言学协会年会论文集（第 1 卷：长篇论文），第 91-101 页 2024 年 8 月 11-16 日 ©2024 计算语言学协会

信息。图 1 展示了一个具有代表性的例子，象征着 "PUBG"（一款电子游戏）如何像一个陷阱一样阻止 "我 "实现 "人生目标"。

为了克服这些挑战，我们希望从 LLMs 中获得洞察力，利用其丰富的世界知识和上下文理解能力来获得图像和文本的深层含义。一种直观高效的方法是利用这些 LLM 生成补充信息，而无需对其进行微调；然后，我们只需微调一个较小的模型，即可在这些信息和隐喻之间建立联系。为了减少 MLLM 的错觉，受 CoT（Wei 等人，2022 年）的启发，我们设计了一种三步法，逐步获取 MLLM 在描述图像、分析文本和整合两种模态信息方面的信息。这种策略的优势如下：首先，它可以为下游模型提供每种模态的附加信息。其次，由浅入深的理解顺序与人类逻辑密切相关，使 LLM 更容易掌握深层含义。此外，后续步骤还可以纠正前面步骤的误解，增强模型的稳健性。

总之，我们利用一种名为 C4MMD 的基于 CoT 的方法来总结 MLLM 中的知识，并通过微调这些知识与隐喻的联系来增强较小模型中的隐喻检测能力。基本思路如图 1 所示：我们首先将图像和文本输入 MLLM，然后获取描述图像、文本及其融合的信息。此外，我们还设计了一个下游模态融合结构，旨在将辅助信息转化为隐喻特征，以实现更准确的分类。具体来说，我们设计了两个辅助任务，重点是确定图像和文本模态中是否存在隐喻。

2 相关工作

早期的隐喻检测任务局限于单一模式，并采用基于规则约束和隐喻词典的方法（Fass，1991；Krishnakumaran和Zhu，2007；Wilks等人，2013）。随着 NLP 领域的蓬勃发展，基于机器学习的方法（Tur-ney 等人，2011 年；Shutova 等人，2016 年）和基于神经网络的方法（Mao 等人，2019 年；Zayed 等人，2020 年）相继出现。这些方法

在引入变压器（Vaswani 等人，2017 年）之后，基于预训练模型的方法逐渐取代了之前的方法，成为当前的主流方法（Cabot 等人，2020 年；Li 等人，2021 年；Lin 等人，2021 年）。Ge 等人（2023 年）将目前的研究工作分为四大类，即附加数据和特征方法（Shutova 等人，2016 年；Gong 等人，2020 年；Kehat 和 Pustejovsky，2021 年）、语义方法（Mao 等人，2019 年；Choi 等人，2021 年；Su 等人，2021 年；Zhang 和 Liu，2022 年；Li 等人，2023b 年；Tian 等人，2023a 年）、基于上下文的方法（Context-based methods，2023b 年）和基于数据的方法（Context-based method，2023a 年）、2023a）、基于语境的方法（Su 等人，2020；Song 等人，2021）和多任务方法（Chen 等人，2020；Le 等人，2020；Mao 等人，2023；Badathala 等人，2023；Zhang 和 Liu，2023；Tian 等人，2023b），其中语义方法和多任务方法已成为近期研究的主要重点。作为一个新兴方向，出现了许多跨图像和文本模态的数据集，这些数据集主要来自社交媒体和广告，产生了大量多语言文本图像模态数据（Zhang 等，2021；Xu 等，2022；Zhang 等，2023a）。与上述从不同模态中提取信息并直接合并的方法不同，我们利用 LLMs，采用 CoT 方法来分析模态之间的特征，从而帮助下游的模态分析。在跨模态融合中的作用

3 方法

我们提出了一种利用 MLLMs 增强隐喻检测的新型框架，称为 C4MMD。我们首先介绍了任务定义（3.1）和完整的模型架构（3.2）。然后，我们利用 CoT 方法（3.3）和下游融合模块的实现（3.4），从 MLLMs 中获取知识。最后，我们简要介绍了训练方法（3.5）。

3.1 任务定义

从形式上看，多模态隐喻检测任务属于典型的多模态分类问题。给定一组跨模态样本对，该任务旨在确定是否存在隐喻特征并提供分类结果。我们的工作重点是检测图像-文本对中的隐喻，因此该任务被表述为：

$$Y = F\left(x^{I}, x^{T}\right) \tag{1}$$

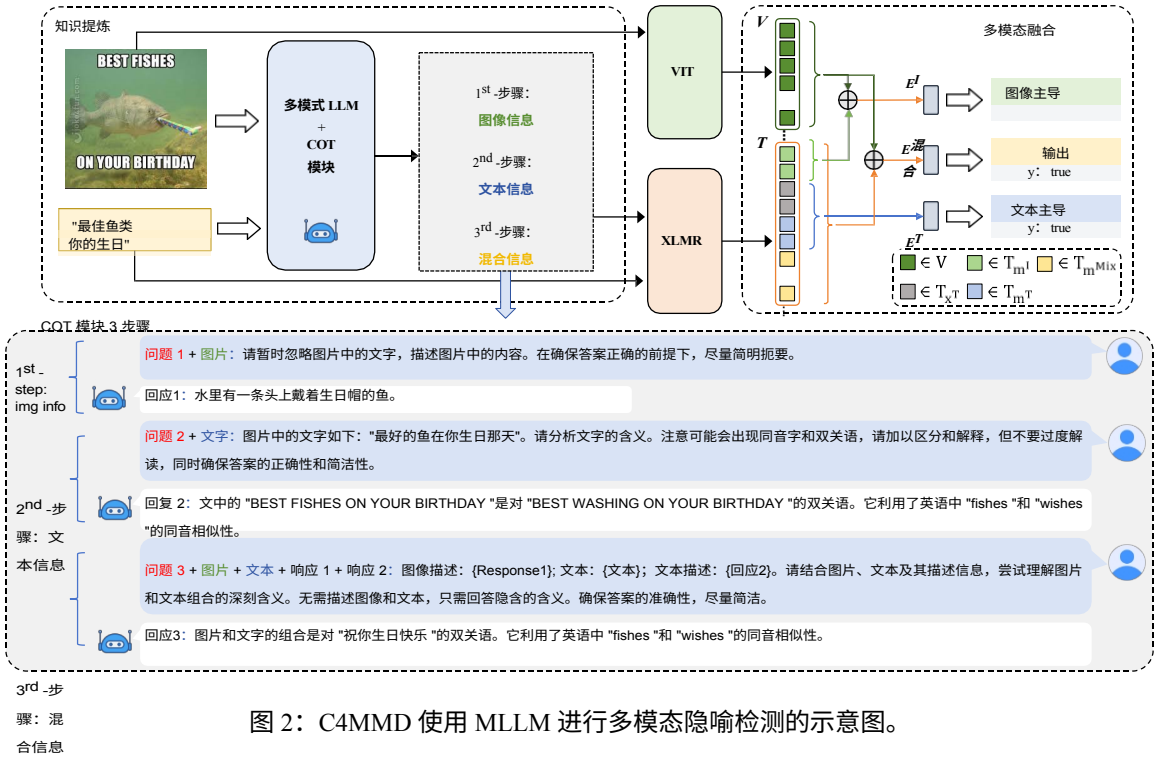


图 2：C4MMD 使用 MLLM 进行多模态隐喻检测的示意图。

其中 x^I 和 x^T 分别表示图像和文本模式的特征。我们的目标是利用更有效的方法 F 来确保分类结果 $\hat{Y}$ 更接近真实值 Y。

3.2 概述

如图 2 所示，C4MMD 的结构由两个主要部分组成：知识汇总模块和多模型融合的下游结构。

在知识总结模块中，我们向 MLLM 提供一对图像-文本，并在 CoT 提示下生成一个三步模板。前两个模板指示 MLLM 专注于一种模式--文本或图像，而忽略另一种模式，以生成解释和见解。在第三步中，MLLM 将两种模式的见解结合起来。根据之前的分析，该模型能够更深入地理解和更全面地整合两种模式。

从 MLLM 获取不同模态的附加文本信息后，我们将其与原始文本合并，形成文本输入。同样，输入图像也被视为视觉模态输入。然后，模型通过特定模态编码器对这些输入进行处理，得出特征向量。

在多模式融合模块中，我们对来自不同模式的矢量进行缩放和组合，并对其进行去矢量化。

开发细粒度分类器。具体来说，我们将补充图像描述向量与视觉模态输入向量相结合作为图像向量，将文本分析向量与文本输入向量相结合作为文本向量，并将这些向量合并形成跨模态向量。然后将这三个向量用于分类目的。分类器使用跨模态向量来检测隐喻，使用图像向量来识别以图像为主的内容，使用文本向量来识别以文本为主的内容。这种方法加强了多模态特征在精确隐喻检测中的应用。

3.3 使用 CoT 方法从 MLLMs 中总结知识

为了引导 MLLM 生成质量更高、信息量更大的特征，我们采用了 CoT 提示。这种方法可以引导 MLLM 跨模态提取更深层次的信息。然后，我们利用这些补充信息来辅助较小的模型实现更好的语义理解和模态融合。总之，我们构建了如下三步提示。

步骤 1.起初，为了确保模型集中精力理解图像中的物体、场景或其他视觉元素(用 x 表示^{I})，而不受文本特征的干扰，我们根据模板 *问题 1* 引导模型理解和解释图像信息：

问题 1：请暂时忽略图片中的文字，描述图片中的内容。在确保答案正确的前提下，尽量简明扼要。

这一步骤可表述如下

$$m^I = MLLM\,(x^I\,,\ \text{问题1}) \tag{2}$$

第 2 步。接下来，为了更好地理解文本中隐藏的含义(由 x 表示^T)，同时排除图像特征的干扰，我们引导模型根据模板*问题 2* 理解和解释文本信息：

问题 2：请分析文本的含义。注意可能会有同音的顺口溜和双关语，请加以区分和解释，但不要过度解读，同时确保答案的正确性和简洁性。

这一步骤可表述如下

$$m^T = MLLM\,(x^T\,,\ \text{Question2}) \tag{3}$$

第 3 步。最终，我们希望模型能综合前两步的结果（用 m^I 和 m^T 表示），并进一步整合图像和文本特征（x^I 和 x^T），从而获得更深刻的跨模态交互信息。我们鼓励模型根据模板*问题 3* 融合不同模态的特征：

问题 3：请结合图片、文字及其描述信息，揣摩图文结合的深刻含义。无需描述图片和文字，只需回答隐含含义。确保答案的准确性，尽量简洁。

这一步骤可表述如下

$$m^{Mix} = MLLM\,(x^I\,,\ x^T\,,\ m^I\,,\ m^T\,,\ \text{Question3}) \tag{4}$$

3.4 隐喻检测的多模态融合

3.4.1 特定模态编码

我们使用一个图像编码器和一个文本编码器来获得图像 x^I 和文本 x^T *的*矢量化编码，以便随后进行模态间融合。考虑到 MLLM 生成的额外信息以文本形式呈现，我们将其视为额外的视觉 m^I、文本 m^T 和混合 m^{Mix} 信息。这些信息与原始文本连接在一起，然后通过文本编码器进行计算处理。

$$\begin{aligned} V &= \text{ViT-Encoder}(x^I)、\\ T &= \text{XLMR-Encoder}(x^T\,,\ m^T\,,\ m^I\,,\ m)^{Mix} \end{aligned} \tag{5}$$

其中，V 是图像编码器的输出，T 是文本编码器的输出。

为了使文本编码器能够在组合过程中区分来自不同模态的文本，我们采用了与 BERT 的段编码类似的方法，为来自每种模态的文本添加额外的可学习参数向量。进入文本编码器的*第 i 个*单词 x_i （$x_i \in \{x^T\,,\ m^T\,,\ m^I\,,\ m^{Mix}\}$）的向量化编码 Emb_i 可表示如下：

$$Emb_i = E_T\,(x_i) + E_P\,(i) + E_S\,(segment(x))_i \tag{6}$$

其中，E_T、E_P 和 E_S 分别代表标记嵌入、位置编码和词段嵌入的可学习矩阵。*词段*（x_i）∈（0，1，2，3）指的是词 x_i 的词段编码，该编码具体由以下公式表示：

$$segment(x_i) = \begin{cases} 1, & \text{如果 } x_i \in m^I \\ 2, & \text{如果 } x_i \in \{x^T\,,\ m^T\} \\ 3, & \text{如果 } x_i \in m^{Mix} \\ 0, & \text{其他} \end{cases} \tag{7}$$

在获得 MLLM 生成的额外模态信息后，我们设计了一种模态融合架构，以促进模态间的融合，并有效利用 MLLM 生成的额外信息来提高隐喻检测能力。

3.4.2 模式融合

在模态融合之前，为了确保两个编码器的向量维度一致，在文字模态中，我们计算所有单词向量的平均值 mean($\boldsymbol{T}$) 作为整个句子的向量表示。对于视觉模式，我们将 CLS 标记的向量 V_{CLS} 作为整个图像的表示。然后，我们使用带有 GeLU 激活函数（Hendrycks 和 Gimpel，2016 年）的线性层将其映射到与文本模态相同的特征空间。计算公式如下

$$V^{reshape} = \mathrm{GeLU}(\boldsymbol{W}\,\boldsymbol{V}_{vCLS} + \boldsymbol{b})_v \qquad (8)$$

考虑到来自不同国家的文本信息

因此，我们将来自两种模态的这两个向量连接起来，得到最终的融合向量表示。这一过程的公式如下

$$E^{Mix} = [V^{reshape}, \text{mean}(T)]。 \quad (9)$$

最后，我们使用一个线性层和一个 softmax 层

。
用于隐喻分类的分类器。

$$\hat{y} = \text{softmax}(W E_{Mix}^{Mix} + b)_{Mix} \quad (10)$$

考虑到隐喻特征来源的多样性，我们采用了两个独立的分类器来对主要由图像模式或文本模式驱动的隐喻进行分类。这样做的目的是在图像和文本融合之前强制检测它们的隐喻特征，从而降低最终分类器的分类复杂度。这种细粒度隐喻检测方法基于以下公式：

$$E^{I} = [V^{reshape}, \text{mean}(T_m \text{ I})] \quad (11)$$

$$E^{T} = \text{mean}([T_x T, T_m \text{ T}]) \quad (12)$$

这里，T_m I、T_x T 和 T_m T 分别代表文本编码向量中描述图像和文本的部分。 最后，两个分类器
用于对文本和图像中的隐喻特征进行分类。这一分类过程的公式如下：

$$\hat{y}^{I} = \text{softmax}(W E_I^I + b)_I \quad (13)$$

$$\hat{y}^{T} = \text{softmax}(W_T E^T + b_T) \quad (14)$$

在上述公式中，W_v，W_{Mix}，W_I 和 W_T 是可训练参数矩阵；b_v，b_{Mix}，b_I 和 b_T 代表偏差矩阵。

3.5 培训

我们的多模态隐喻检测模型的训练目标涉及三个不同损失函数的整合，分别表示为 L_I 、L_T 和 L_M。损失函数如下

{L_I ， L_T， L_M }， ={ŷ,ŷ^T } 和 Yrep-

其中 Ŷ

|D ||ME

根据模型的预测结果和真实值，L_{CE} 是交叉熵损失函数。为了优化整体性能，我们将总损失 L_{sum} 定义为这些单个损失的加权组合。最终的损失函数表述为

$$l_{sum} = 0.5 \cdot l_I + 0.5 \cdot l_T + l_M \quad (16)$$

4 实验

在本节中，我们首先介绍用于验证我们方法的数据集以及实验设置。随后，我们将报告实验结果，并对这些结果进行分析。

4.1 数据和环境

我们选择了 Xu 等人（2022 年）建立的多模态隐喻数据集，该数据集由从社交媒体上收集的 10,000 张 meme 图像组成。我们使用 OCR 方法从这些图片中提取文本信息，构建了多模态隐喻数据集，其中包括 6000 个中文词条和 4000 个英文词条。除了隐喻的分类标签外，他们还标注了隐喻的来源及其相关情绪。

$$L = \frac{1}{|D|_{ME}} \sum_{i=1} L_{CE}(\hat{Y}, Y) \quad (15)$$

其中，D_{ME} 是数据集中的样本数，公式的参数为 L = D 。

所有训练模型的学习率设定为 1e-5，批量大小为 8，训练时间为 100 个历元，并设置了提前停止机制。数据集被随机洗牌，并按 6:2:2 的比例分为训练集、验证集和测试集。所有实验均在单个 3090- 24G GPU 上进行。我们方法的最终结果是取五个不同运行种子的平均值，平均单次运行时间为 20-30 分钟。最后，根据 F1 分数评估了模型的性能。

在对 LLM 进行微调时，采用了 Low-Rank Adaptation (LoRA Hu 等人 (2021))微调方法。所有设置都沿用了 Alpaca-LoRA* 中的设置。

4.2 基准方法 语言模型

我们针对这项任务测试了几种常见的预训练模型，包括 AutoEncoder M-BERT（Pires等人，2019年）、XLM-R（Conneau等人，2019年）、

* Alpaca-LoRA

模式	型号	模型	ACC	P.	R.	F1.
语言模式	自动编码器	M-BERT-base	74.60	61.25	76.93	68.20
		XLM-R-base	83.32	78.57	72.71	75.53
	自动回归模型	M-T5-基座	83.86	80.25	71.91	75.85
		大号 M-BART	83.52	78.79	73.14	75.86
	法学硕士	LLaMA2-7b (LoRA)	83.07	78.23	72.29	75.15
		ChatGLM3-6b (LoRA)	84.81	82.22	72.86	77.26
愿景模型	CNN 模型	ResNet50	75.25	69.53	53.59	60.52
		VGG16	77.69	72.48	59.63	65.43
		ConvNeXt-base	79.33	74.75	62.87	68.30
	变压器型号	ViT 基础	74.75	65.50	60.62	62.97
		斯温变压器基地	78.83	77.82	56.26	65.31
多模式模型		VILT	83.13	78.01	72.86	75.35
		InternLM-XComposer-7b （零拍摄）	67.50	30.83	17.29	22.16
		BLIP2-2.7b（零镜头）	38.33	33.44	82.97	47.05
		BLIP2-2.7b (LoRA)	85.66	80.61	78.34	79.46
相关工作		CLIP（Zhao 等人，2023 年）	75.05	60.83	**83.07**	70.23
		维里奥（缪尼戈夫，2020 年）	84.30	79.97	79.97	76.74
		酷网（Xiao 等人，2023 年）	77.49	66.84	72.29	69.46
		MultiCMET （Zhang 等人，2023b）	85.66	82.69	75.25	78.79
C4MMD（我们的）			**87.70**	**83.33**	81.58	**82.44**

表 1：不同方法在多模态隐喻检测任务中的结果。

2019），以及自动回归模型 M- T5（薛等人，2020）和 M-BART（刘等人，2020）。此外，我们还使用 LLaMA2（Touvron 等人，2023 年）和 ChatGLM3（Zeng 等人，2022 年）评估了 LLM 在这项任务中的能力，因为它们在中文和英文语料库中都有很好的表现。我们使用 LoRA 分别对这两个模型进行了微调。

视觉模型

我们还测试了视觉领域的模型，包括卷积神经网络（CNN）模型，如 VGG（Simonyan 和 Zisser- man，2014 年）、ResNet（He 等人，2016 年）和 Con- vNeXt（Liu 等人，2022 年），以及基于变形器架构的模型，如 ViT（Dosovit- skiy 等人，2020 年）和 Swin Transformer（Liu 等人，2021 年）。

多模式模型

在多模式模型领域，我们选择了 VILT（Kim 等人，2021 年）、BLIP2（Li 等人，2023a 年）、和 InternLM-XComposer（Zhang 等人，2023c），以测试它们处理隐喻检测任务的能力。这三个模型都采用了 Trans- former 架构，但它们在模型大小上有很大不同。我们测试了这些 MLLM 在零镜头设置和 LoRA 微调下的能力。

其他相关作品

我们还探讨了与我们的任务相关的其他作品，从而使我们的比较分析更具可信度。下面，我们将详细介绍这些作品。

- **CLIP**（Zhao 等人，2023 年）：针对仇恨备忘录检测任务的多种模型评估。我们采用性能最佳的 CLIP 来评估其在多模态隐喻检测任务中的有效性。
- **维里奥**（Muennighoff，2020 年）：这是一种优秀的方法，在 "憎恨备忘录挑战赛"（Hate-ful Memes Challenge）中获得第二名。它使用 OCR 和实体识别技术从备忘录中提取文本和视觉特征，以更好地完成备忘录有害性检测任务。
- **CoolNet**（肖等人，2023 年）：提取文本句法结构，提高模型对 Twitter 多模态数据的情感分析能力。
- **MultiCMET**（Zhang 等人，2023b）：中国多模态隐喻检测任务的基础模型。它使用 CLIP 模型来生成附加信息，以协助模态之间的融合。

4.3 主要成果

表 1 显示了不同模型在多模态隐喻检测任务中的能力。这里我们只评估了主要的分类结果 $y^$ 。

模型	ACC	P.	R.	F1.
我们的	87.70	83.33	81.58	82.44
-融合模式	85.66	77.87	83.12	80.41
-CoT功能	85.06	78.42	79.75	79.08
-视觉编码器	86.25	78.36	84.53	81.33

表 2：隐喻检测模型中各组成部分的消减研究。

虚拟机	LM	ACC	P.	R.	F1.
ResNet	M-BERT	82.38	78.29	69.48	73.62
VGG		85.86	84.60	73.42	78.61
ViT		85.75	81.73	76.99	79.27
ViT	M-T5	76.66	68.51	62.64	65.44
	M-BART	80.21	70.97	75.14	72.92
	XLMR	86.39	83.68	76.54	79.92

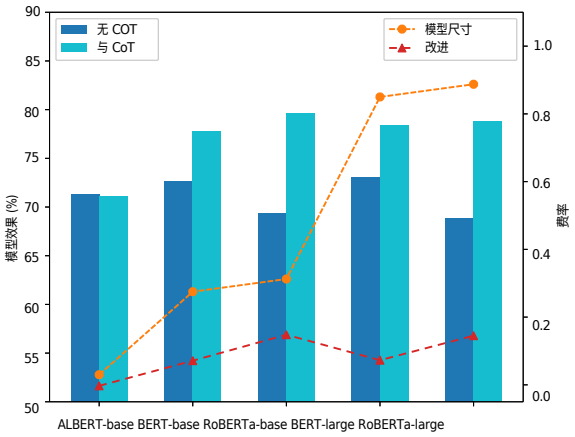
表 3：不同语言和视觉模型组合对隐喻检测任务的影响，VM 代表视觉模型，LM 代表语言模型。然后，我们使用线性层来融合两种模式的特征。

我们没有评估 $\hat{y}^I$ 和 $\hat{y}^T$ 这两项子任务的结果，因为这两项子任务主要是为主要任务服务的。

我们的方法在中文和英文样本集上都取得了最佳结果。考虑到 MLLM（InternLM-XComposer-7b）直接生成的结果，我们允许它直接生成图像和文本的附加特征，有效地利用了大型模型的能力。这种方法与下游分类器相结合，产生了叠加效应。

多模态模型的性能差异很大，大多数模型都没有超过语言模型。这凸显了文本模态在识别多模态隐喻方面的重要性。多模态隐喻模型在零镜头场景中表现不佳，部分原因是我们设计了提示模板。不过，主要原因是模型无法理解任务。令人鼓舞的是，在对 BLIP2 进行微调后，它的能力超过了所有其他比较方法。这证明了在任务中图像和文本模式之间的交互是有益的，也证明了大型模型在经过微调后可以有效地理解和处理这项任务。

在相关工作中，与我们的研究密切相关的研究，如 Zhang 等人（2023b）和 Muennighoff（



2020）的研究，都取得了不俗的成绩。然而，肖等人（2023）的 Twitter 情感分类与我们的任务有些不同，因此表现较弱。

图 3：生成或不生成 CoT 的不同大小模型的效果和改进率。我们将模型大小的截距控制在 0-1 之间，以显示单个数字的改进效果。

4.4 不同因素的影响

表 2 显示了我们的模型在进行消融实验后所展示的效果。

用线性层取代模型中的融合结构会导致性能显著下降。这表明有必要增加融合结构，以帮助模型理解 MLLM 生成的额外特征。此外，取消 MLLM 的 CoT 生成方法，仅依靠一步生成方法，会导致更明显的性能下降。这也说明 CoT 方法可以生成更好的附加特征，从而帮助下游模型做出更准确的判断。

有趣的是，当我们移除图像处理模块时，模型的性能仅略有下降。这表明，MLLM 可以为较小的模型提供一定程度的视觉信息，但更全面的模型形成仍需要视觉模型的贡献。

4.5 不同语言视觉模型组合的影响

在模态融合过程中，我们测试了多种视觉和文本模型的能力。在测试视觉模型时，语言模型统一设置为 MBERT，而在测试语言模型时，则统一使用 ViT。

从表 3 和表 1 的数据来看，虽然在单一模式设置中，视觉模型 VGG 和文本模型 M-T5 的每

图 4：案例研究实例。

在模态融合方面，ViT 和 XLM-R 的组合表现优于其他所有组合。

ResNet + MBERT 和 VGG + MBERT 的组合也是 Met-Meme 提出的基准模型（Xu 等人，2022 年）。根据结果，我们报告了与他们相同的结果。

4.6 语言模型大小的影响

图 3 展示了不同规模的模型在我们的架构下的能力。考虑到改进率通常介于 0 和 1 之间，而模型大小通常在数亿之间，我们将所有模型大小除以 4 亿，使其比例介于 0 和 1 之间，这样就可 以在同一张图上同时显示模型大小和改进率。很明显，随着模型大小的增加，特别是当模型最初较小时，性能会逐渐得到明显改善。当模型太小时，额外的文本信息并不会产生积极的效果，反而有可能对模型的性能产生负面影响。只有当模型规模增大时，模型才有能力理解更长的上下文信息。

4.7 案例研究

为了进一步探索我们所提模型的有效性，我们从图 4 所示的测试数据集中选取了两个示例。

第一个示例演示了图像引导的

比喻。通过直接比较海豹和土豆，它描述了看太多可爱海豹的后果。通过对图像的理解，MLLM 准确地识别出了海豹和土豆之间的相似之处，从而帮助下游模型做出了正确的判断。

在第二个例子中，MLLM 从图像和文本中识别出特征，然后将这些特征结合起来，正确理解了备忘录所表达的幽默含义。下游模型准确地识别出它不包含隐喻特征。相比之下，缺乏大模型附加信息的方法仅根据 "像个淑女 "这一短语就判断其具有隐喻性，从而导致错误判断。

5 结论

我们的研究旨在利用先进的 MLLM 解决多模态隐喻解释的难题。我们设计了一种三步法，利用 CoT 提示从图像和文本中提取更丰富的信息。事实证明，来自 MLLMs 的增强知识对于增强较小模型以掌握每种模态中的隐喻特征以及模态融合至关重要。这项工作不仅推进了多模态隐喻检测，还为未来研究探索 MLLMs 在解决复杂语言和视觉挑战方面的潜力铺平了道路。

局限性

我们认为，我们工作的主要局限性在于只在多语言 meme 数据集中测试了我们的隐喻检测能力，而没有扩展到 meme 数据集中的其他子任务，如有害性检测，也没有扩展到其他多模态数据集中的隐喻检测。不过，尽管缺乏实验数据，我们对我们的工作在这些方向上的适用性充满信心，这也将是我们未来的研究重点之一。

此外，关于 meme 数据集，我们没有找到使用许可证，也没有过滤数据中潜在的有害性或冒犯性，包括 MLLM 生成的额外特征，其中可能包含有毒数据，因此预先发送了冒犯性和有害性风险。

虽然我们对自己的模型采用了五次测试取平均值的方法，但对于其他比较方法，我们只是取第一次测试的结果列入表格。我们承认这可能会带来一些误差，但我们相信，即使用同样的方法对比较方法进行测试，我们的方法仍然会表现出压倒性的优越性。

参考资料

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang.2022.铃声响起：用于视频中多模态隐喻检测的语料库和方法。*arXiv 预印本 arXiv:2301.01134*.

Naveen Badathala、Abisek Rajakumar Kalarani、Tejpalsingh Siledar 和 Pushpak Bhattacharyya。2023.天作之合：用于夸张和隐喻检测的多任务框架。*ArXiv 预印本 arXiv:2305.17480*.

Pere-Lluís Huguet Cabot、Verna Dankers、David Abadi、Agneta Fischer 和 Ekaterina Shutova。2020.政治背后的语用学：政治话语中的隐喻、框架和情感建模。In *Find-ings of the Association for computational linguistics: emnlp 2020*, pages 4479-4488.

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov.2020.基于多任务转换器的成语隐喻去保护架构：2020 年隐喻共享任务中的 Ets 团队。*第二届比喻语言处理研讨会论文集*》，第 235-243 页。

Minjin Choi、Sunkyung Lee、Eunseong Choi、Heesoo Park、Junhyuk Lee、Dongwon Lee 和 Jongwuk Lee。2021.Melbert: Metaphor detection via contextual- ized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov.2019.大规模无监督跨语言表征学习。*ArXiv 预印本 arXiv:1911.02116*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.一幅图像胜过 16x16 个单词：规模图像识别变换器。*arXiv 预印本 arXiv:2010.11929*.

Dan Fass.1991年：计算机辨别隐喻和转喻的方法。*计算语言学*》，17（1）：49-90。

Mengshi Ge, Rui Mao, and Erik Cambria.2023.计算隐喻处理技术概览：从识别、解释、生成到应用.*人工智能评论*》，第 1-67 页。

Aniruddha Ghosh and Tony Veale.2016.使用神经网络进行冷嘲热讽。*第七届主观性、情感和社交媒体分析计算方法研讨会论文集*》，第 161-169 页，加利福尼亚州圣迭戈。计算语言学协会。

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat.2020.Illinimet：利用上下文和语言信息进行隐喻检测的 Illinois 系统。*第二届比喻语言处理研讨会论文集*》，第 146-153 页。

何开明、张翔宇、任绍清、孙健。2016.图像识别的深度残差学习。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770- 778.

Dan Hendrycks and Kevin Gimpel.2016.Gaus- sian error linear units (gelus). *ArXiv preprint arXiv:1606.08415*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.2021.Lora： *ArXiv preprint arXiv:2106.09685*.

Gitit Kehat 和 James Pustejovsky。2021.使用可见性嵌入的神经隐喻检测。*SEM 2021 论文集：第十届词汇与计算语义学联合会议*》，第 222-228 页。

Wonjae Kim, Bokyung Son, and Ildoo Kim.2021.Vilt：无需卷积或区域监督的视觉语言转换器。*国际机器学习大会*，第 5583-5594 页。PMLR.

Saisuresh Krishnakumaran 和 Xiaojin Zhu.2007.利用词汇资源狩猎难以捉摸的隐喻。*比喻语言计算方法研讨会论文集*》，第 13-20 页。

Duong Le, My Thai, and Thien Nguyen.2020.利用图神经网络和词义辨析进行隐喻检测的多任务学习。In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8139-8146.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi.2023a.Blip-2：使用冻结图像编码器和大型语言模型进行语言图像预训练*的*引导。

李树群、杨亮、何卫东、张诗琦、曾静洁和林鸿飞。2021.用于序列隐喻识别的标签增强型分层语境化表示法。*自然语言处理实证方法 2021 年会议论文集*》，第 3533-3543 页。

李玉成、王顺、林成华、弗兰克-盖林。2023b.通过外显基本含义建模进行隐喻检测。*arXiv 预印本 arXiv:2305.17268*.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen.2021.Cate：半监督学习的隐喻检测对比预训练模型。*自然语言处理经验方法 2021 年会议论文集*》，第 3888-3898 页。

刘银汉、顾嘉涛、纳曼-戈亚尔、李娴、谢尔盖-埃杜诺夫、马尔扬-加兹维尼内贾德、迈克-刘易斯和卢克-泽特勒莫耶。2020.神经机器翻译的多语言去噪预训练。*计算语言学协会论文*集》，8：726-742。

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.2021.Swin变换器：使用移位窗口的分层视觉变换器。*IEEE/CVF 计算机视觉国际会议论文集*》，第 10012-10022 页。

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie.2022. 面向 2020 年代的 convnet。*IEEE/CVF 计算机视觉与图像识别会议论文集*》，第 11976-11986 页。

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria.2023.在线隐喻：计算隐喻处理在线系统。*第61届计算语言学协会年会论文集（第3卷：系统演示）*，第127-135页。

毛锐、林成华和弗兰克-盖林。2019.受语言学理论启发的端到端序列隐喻识别。*计算语言学协会第 57 届年会论文集*》，第 3888-3898 页。

Saif Mohammad、Ekaterina Shutova 和 Peter Turney 。2016.隐喻作为情感的媒介：实证研究。In *Proceedings of the Fifth Joint Confer- ence on Lexical and Computational Semantics*, pages 23-33.

Niklas Muennighoff.2020.Vilio：最先进的视觉语言学模型应用于仇恨备忘录。

Yair Neuman、Dan Assaf、Yohai Cohen、Mark Last、Shlomo Argamon、Newton Howard 和 Ophir Frieder。2013.大型文本语料库中的隐喻识别。*PloS one*, 8(4):e62343.

Telmo Pires、Eva Schlinger 和 Dan Garrette。2019.*ArXiv preprint arXiv:1906.01502*.

Ekaterina Shutova、Douwe Kiela 和 Jean Maillard。2016.黑洞和白兔：用视觉特征识别隐喻。*计算语言学协会北美分会 2016 年会议论文集：人类语言技术*》，第 160-170 页。

Karen Simonyan and Andrew Zisserman.2014.用于大规模图像识别的深度卷积网络》，*arXiv preprint arXiv:1409.1556*.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu.2021.通过上下文关系学习的动词隐喻检测。*计算语言学协会第 59 届年会暨第 11 届自然语言处理国际联合会议论文集》（第 1 卷：长篇论文）*，第 4240-4251 页。

Chang Su, Kechun Wu, and Yijiang Chen.2021.基于语言学理论的外部知识融入增强隐喻检测。In *Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021*，第 1280-1287 页。

苏传东、Fumiyo Fukumoto、黄晓曦、李继义、王荣波、陈志群。2020.Deepmet：用于标记级隐喻检测的阅读理解范式。*第二届比喻语言处理研讨会论文集*》，第 30-39 页。

Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng.2023a.隐喻检测中的概念属性相似性和领域不一致性建模。*自然语言处理经验方法 2023 年会议论文集*》，第 7736-7752 页。

Yuan Tian、Nan Xu、Wenji Mao 和 Daniel Zeng.2023b.隐喻检测中的概念属性相似性和领域不一致性建模。*自然语言处理经验方法 2023 年会议论文集*》，第 7736-7752 页，新加坡。计算语言学协会。

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al.Llama 2: Open founda- tion and fine-tuned chat models. *ArXiv preprint arXiv:2307.09288*.

Peter Turney、Yair Neuman、Dan Assaf 和 Yohai Cohen.2011.通过具体和抽象语境识别字面意义和隐喻意义。*2011年自然语言处理实证方法大会论文集*》，第680-690页。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin.2017.注意力就是你所需要的一切。*神经信息处理系统进展*》，第30期。

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al.大型语言模型中的思维链提示引发再学习。*神经信息处理系统进展*》，35:24824-24837。

Yorick Wilks、Adam Dalton、James Allen 和 Lucian Galescu。2013.使用大规模词汇资源和传统隐喻提取进行隐喻自动检测。*首届 NLP 隐喻研讨会论文集*》，第 36-44 页。

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He.2023.基于多模态方面情感分析的跨模态细粒度配准与融合网络.*信息处理与管理*，60（6）：103508.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia.2022.Met-meme：富含隐喻的多模态 meme 数据集。*第 45 届 ACM SIGIR 信息检索研究与发展国际会议论文集*》，第 2887-2899 页。

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel.mt5: A massive multilingual pre-trained text-to-text transformer. *ArXiv preprint arXiv:2010.11934*.

Omnia Zayed、John P McCrae 和 Paul Buitelaar。2020.关系级隐喻识别的语境调制。*arXiv preprint arXiv:2010.05633*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al.Glm-130b: An open bilingual pre-trained model. *ArXiv preprint arXiv:2210.02414*.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin.2023a.Multicmet：用于理解多模态隐喻的新型中文基准。In *Findings of the Association for Computational Linguistics：EMNLP 2023*，第 6141-6154 页。

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin.2023b.Multicmet：用于理解多模态隐喻的新型中文基准。*计算语言学协会论文集：EMNLP 2023*，第 6141-6154 页。

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin.2021.Multimet：用于隐喻理解的多模态数据集。In *Pro- ceedings of the 59th Annual Meeting of the Asso- ciation for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214- 3225.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al.Internlm-xcomposer：用于高级文本图像理解和合成的视觉语言大模型。*arXiv 预印本 arXiv:2309.15112*.

Shenglong Zhang and Ying Liu.2022.通过语言学增强连体网络的隐喻保护。*第 29 届国际计算语言学大会论文集》*，第 4149-4159 页。

Shenglong Zhang and Ying Liu.2023.用于端到端隐喻检测的对抗性多任务学习。*ArXiv 预印本 arXiv:2305.16638*.

Bryan Zhao, Andrew Zhang, Blake Watson, Gillian Kearney, and Isaac Dale.2023.视觉语言模型及其在 "仇恨备忘录挑战 "中的表现综述。*arXiv预印本arXiv:2305.06159*.