Exploring Chain-of-Thought for Multi-modal Metaphor Detection

Yanzhi Xu[*], **Yueying Hua**[*], **Shichen Li** and **Zhongqing Wang**[†]
Natural Language Processing Lab, Soochow University, Suzhou, China
{yzxuxyz,yyhua1224}@stu.suda.edu.cn
scli_21@outlook.com, wangzq@suda.edu.cn

Abstract

Metaphors are commonly found in advertising and internet memes. However, the free form of internet memes often leads to a lack of high-quality textual data. Metaphor detection demands a deep interpretation of both textual and visual elements, requiring extensive common-sense knowledge, which poses a challenge to language models. To address these challenges, we propose a compact framework called C4MMD, which utilizes a **C**hain-of-Thought(CoT) method **for** **M**ulti-modal **M**etaphor **D**etection. Specifically, our approach designs a three-step process inspired by CoT that extracts and integrates knowledge from Multi-modal Large Language Models(MLLMs) into smaller ones. We also developed a modality fusion architecture to transform knowledge from large models into metaphor features, supplemented by auxiliary tasks to improve model performance. Experimental results on the MET-MEME dataset demonstrate that our method not only effectively enhances the metaphor detection capabilities of small models but also outperforms existing models. To our knowledge, this is the first systematic study leveraging MLLMs in metaphor detection tasks. The code for our method is publicly available at https://github.com/xyz189411yt/C4MMD.

1 Introduction

Metaphors are highly prevalent in our everyday expressions and writings, which can have a range of impacts on downstream tasks in Natural Language Processing (NLP), such as semantic understanding (Neuman et al., 2013), sentiment analysis(Ghosh and Veale, 2016; Mohammad et al., 2016) and other tasks. In recent years, the rise of social media has sparked interest in multi-modal metaphors. As a result, several datasets for multi-modal metaphors have been proposed (Zhang et al., 2021, 2023a; Alnajjar et al., 2022).

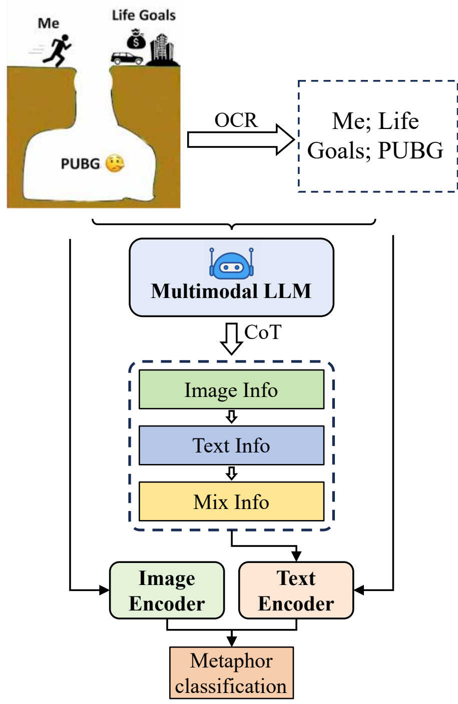

Figure 1: An example of multi-modal metaphor detection.

Current research on multi-modal metaphor detection is still in its early stages. The primary challenge lies in the complexity and variety of multi-modal metaphors. Compared to single-modality detection, multi-modal metaphor detection not only spots metaphors in sentences but also categorizes them as image-dominated, text-dominated, or complementary. The second major challenge arises from the poor quality of textual content, mainly sourced from advertisements and memes on social media. Texts give the image more metaphorical features. Recent efforts use OCR (Optical Character Recognition) to extract texts in the image. However, only relying on OCR to convert them into parallel texts leads to the loss of texts' positional

[*]Equal contribution
[†]Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 91–101
August 11-16, 2024 ©2024 Association for Computational Linguistics

information. Figure 1 presents a representative example, symbolizing how 'PUBG' (a video game) acts like a trap preventing "me" from achieving my "life goals".

To overcome these challenges, we hope to gain insights from LLMs, utilizing their rich world knowledge and contextual understanding capabilities to obtain deeper meanings of both images and text. An intuitive and efficient approach is to use these LLMs to generate supplementary information without fine-tuning them; we then only need to fine-tune a smaller model to establish connections between this information and metaphors. To reduce the illusion of MLLMs, inspired by CoT (Wei et al., 2022), we have designed a three-step method that progressively acquires the MLLM's information in describing images, analyzing text, and integrating information from both modalities. The advantages of this strategy are as follows: First, it can provide downstream models with additional information for each modality. Second, the shallow-to-deep understanding sequence aligns closely with human logic, making it easier for the LLM to grasp deeper meanings. Furthermore, subsequent steps can correct misunderstandings from earlier steps, enhancing the model's robustness.

Overall, we utilize a CoT-based method called C4MMD to summarize knowledge from MLLMs and enhance metaphor detection in smaller models by fine-tuning them to link this knowledge with metaphors. The basic idea is shown in Figure 1, we first input images and text into the MLLM and obtain information describing the image, text, and their fusion. Furthermore, we have designed a downstream modality fusion structure, which is intended to translate supplementary information into metaphorical features for more accurate classification. Specifically, we have designed two auxiliary tasks focused on determining the presence of metaphors within the image and text modalities.

2 Related Work

Early metaphor detection tasks were confined to a single modality and employed methods based on rule constraints and metaphor dictionaries (Fass, 1991; Krishnakumaran and Zhu, 2007; Wilks et al., 2013). With the flourishing development in the field of NLP, machine learning-based methods (Turney et al., 2011; Shutova et al., 2016) and neural network-based methods (Mao et al., 2019; Zayed et al., 2020) have successively emerged. Following the introduction of the Transformer (Vaswani et al., 2017), methods based on pre-trained models gradually supplanted the former methods and became the current mainstream approach (Cabot et al., 2020; Li et al., 2021; Lin et al., 2021). Ge et al. (2023) have categorized current efforts into four main directions, namely additional data and feature methods (Shutova et al., 2016; Gong et al., 2020; Kehat and Pustejovsky, 2021), semantic methods (Mao et al., 2019; Choi et al., 2021; Su et al., 2021; Zhang and Liu, 2022; Li et al., 2023b; Tian et al., 2023a), context-based methods (Su et al., 2020; Song et al., 2021), and multitask methods (Chen et al., 2020; Le et al., 2020; Mao et al., 2023; Badathala et al., 2023; Zhang and Liu, 2023; Tian et al., 2023b), where semantic methods and multitask methods have become the primary focus of recent research.

As an emerging direction, numerous datasets across image and text modalities have emerged, primarily sourced from social media and advertisements, yielding extensive multilingual text-image modal data (Zhang et al., 2021; Xu et al., 2022; Zhang et al., 2023a). Unlike the aforementioned approaches that extract information from different modalities and directly merge them, we leverage LLMs employing the CoT method to analyze features between modalities, aiding downstream models in cross-modal fusion.

3 Method

We propose a novel framework called C4MMD using MLLMs to enhance metaphor detection. We first introduce the task definition(3.1) and the complete model architecture((3.2). After that, we elaborate on knowledge acquisition from MLLMs using the CoT method(3.3) and the implementation of the downstream fusion module(3.4). Finally, we provide a brief exposition of the training methodology (3.5).

3.1 Task Definition

Formally, the task of multi-modal metaphor detection falls under the typical category of multi-modal classification problems. Given a set of cross-modal sample pairs, the task aims to determine whether metaphorical features are present and provide a classification result. Our work focuses on the detection of metaphors in image-text pairs, thus the task is represented as:

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

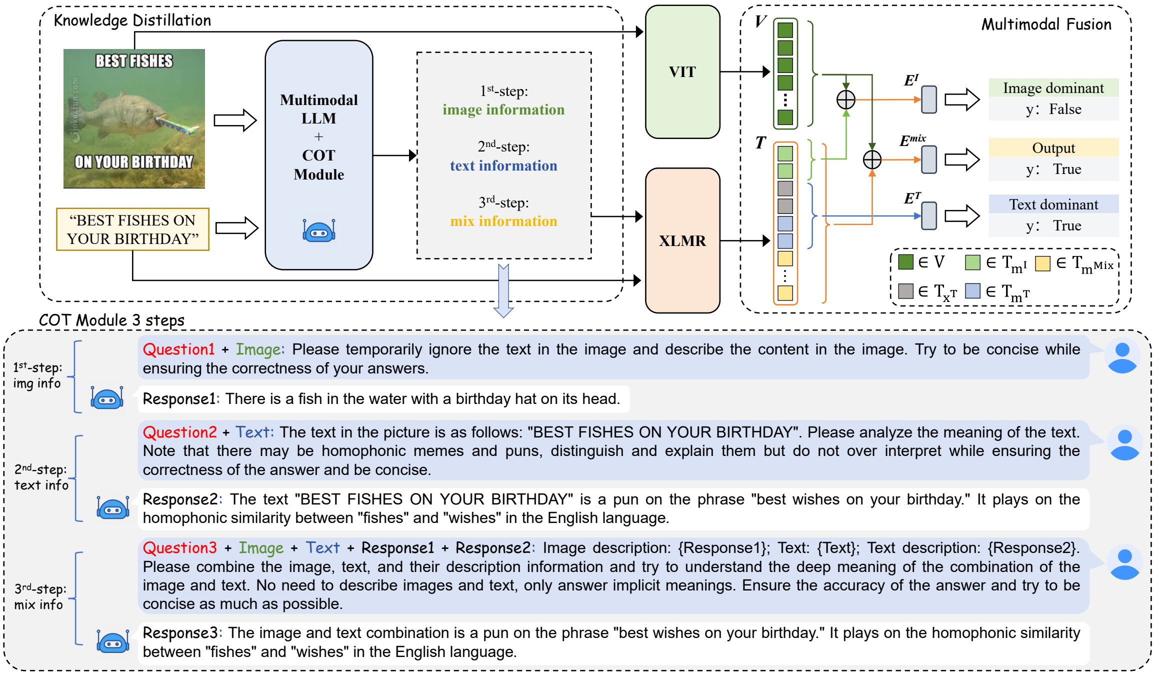

Figure 2: An illustration of C4MMD using the MLLM for multi-modal metaphor detection.

where x^I and x^T respectively denote the features of the image and text modalities. Our objective is to utilize a more effective method F to ensure that the classification result $\hat{Y}$ more closely aligns with the true value Y.

3.2 Overview

As shown in Figure2, the architecture of C4MMD consists of two primary components: a knowledge summarization module and a downstream structure for multi-model fusion.

In the knowledge summarization module, we provide an image-text pair to the MLLM and design a three-step template with CoT prompting. The first two templates instruct the MLLM to focus exclusively on a single modality—either text or image, ignoring the other to generate explanations and insights. In the third step, the MLLM combines insights from both modalities. Based on previous analyses, the model achieves a deeper understanding and a fuller integration of both modalities.

After obtaining additional textual information for different modalities from the MLLM, we merge this with the original texts to form a textual input. Similarly, the input image is treated as the visual modality input. The model then processes these inputs through modality-specific encoders to derive feature vectors.

In the multi-model fusion module, we scale and combine vectors from different modalities and develop a fine-grained classifier. Specifically, we integrate the supplementary image description vector with the visual modality input vector as the image vector, combine the text analysis vector with the textual input vector as the text vector, and merge these to form a cross-modal vector. These three vectors are then used for classification purposes. The classifier uses the cross-modal vector to detect metaphors, the image vector to identify image-dominated content, and the text vector for text-dominated content. This approach enhances the use of multi-modal features for precise metaphor detection.

3.3 Knowledge Summarization from MLLMs Using the CoT Method

To guide the MLLM in generating higher-quality and more informative features, we employ CoT prompting. This method directs the MLLMs to extract deeper information across modalities. We then utilize this supplementary information to assist the smaller model in achieving better semantic understanding and modality fusion. In conclusion, we construct the three-step prompts as follows.

STEP1. Initially, to ensure that the model concentrates on comprehending objects, scenes, or other visual elements in the image(Represented by x^I) without interference from textual features, we guide the model to understand and interpret the image information based on a template $Question1$:

> Question1: Please temporarily ignore the text in the image and describe the content in the image. Try to be concise while ensuring the correctness of your answers.

This step can be formulated as follows:

$$m^I = MLLM(x^I, \text{Question1}) \tag{2}$$

STEP2. Next, to better comprehend the hidden meanings in the text(Represented by x^T) while excluding any interference from image features, we guide the model to understand and interpret the textual information according to a template $Question2$:

> Question2: Please analyze the meaning of the text. Note that there may be homophonic memes and puns, distinguish and explain them but do not over interpret while ensuring the correctness of the answer and be concise.

This step can be formulated as follows:

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

STEP3. Ultimately, we aspire for the model to synthesize the results from the previous two steps(Represented by m^I and m^T) and further integrate the image and text features(x^I and x^T), thereby obtaining more profound cross-modal interaction information. We encourage the model to fuse features from different modalities according to template $Question3$:

> Question3: Please combine the image, text, and their description information and try to understand the deep meaning of the combination of the image and text. No need to describe images and text, only answer implicit meanings. Ensure the accuracy of the answer and try to be concise as much as possible.

This step can be formulated as follows:

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

3.4 Multi-modal Fusion for Metaphor Detection

After obtaining additional modal information generated by the MLLM, we designed a modal fusion architecture to facilitate inter-modal integration and effectively leverage the extra information produced by the MLLM to enhance metaphor detection capabilities.

3.4.1 Modality-Specific Encoding

We use an image encoder and a text encoder to obtain vectorized encodings of the image x^I and text x^T for subsequent inter-modal fusion. Considering the additional information generated by the MLLM is presented in text form, we treat it as extra visual m^I, textual m^T, and mixed m^{Mix} information. This information is concatenated with the original text and then processed through the text encoder for computation.

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

where V is the output of the image encoder, and T is the output of the text encoder.

To enable the text encoder to distinguish between texts from different modalities during computation, we adopt a method similar to BERT's segment encoding by adding extra learnable parameter vectors for the text from each modality. The vectorized encoding Emb_i of the i-th word x_i ($x_i \in \{x^T, m^T, m^I, m^{Mix}\}$) entering the text encoder can be represented as follows:

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

where E_T, E_P and E_S represent learnable matrices for token embeddings, positional encodings, and segment embeddings, respectively. The term $segment(x_i) \in (0, 1, 2, 3)$ refers to the segment encoding of the word x_i, this encoding is specifically represented by the following formula:

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

3.4.2 Modality Fusion

Before modal fusion, to ensure the vector dimensions from both encoders are consistent, in the textual modality, we compute the average of all word vectors mean($\boldsymbol{T}$) as the vector representation of the entire sentence. For the visual modality, we take the vector of the CLS token V_{CLS} as the representation of the entire image. Then, we use a linear layer with a GeLU activation function (Hendrycks and Gimpel, 2016) to map it to the same feature space as the textual modality. The formula is represented as follows:

$$V^{reshape} = \text{GeLU}(\boldsymbol{W_v V_{CLS}} + \boldsymbol{b_v}) \tag{8}$$

Considering that the text information from different modalities generated by the large model has already undergone a degree of fusion within the text encoder, we therefore concatenate these two vectors from both modalities to obtain the final fused vector representation. The formula for this process is as follows:

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

Finally, we use a linear layer and a softmax classifier for metaphor classification.

$$\hat{y} = \text{softmax}(\boldsymbol{W_{Mix}}\boldsymbol{E^{Mix}} + \boldsymbol{b_{Mix}}) \tag{10}$$

Considering the diverse sources of metaphorical features, we employ two separate classifiers to categorize metaphors predominantly driven by either the image modality or the text modality. The aim is to force the detection of metaphorical features in both image and text before their fusion, thereby reducing the classification complexity for the final classifier. This approach of fine-grained metaphor detection is based on the following formula:

$$\boldsymbol{E^{I}} = [V^{reshape}, \text{mean}(\boldsymbol{T_{m^I}})] \tag{11}$$

$$\boldsymbol{E^{T}} = \text{mean}([\boldsymbol{T_{x^T}}, \boldsymbol{T_{m^T}}]) \tag{12}$$

Here, $\boldsymbol{T_{m^I}}$, $\boldsymbol{T_{x^T}}$ and $\boldsymbol{T_{m^T}}$ respectively represent the parts of the text encoding vector that describe the image and the text. Finally, two classifiers are used to categorize the metaphorical features in the text and the image. The formula for this classification process is as follows:

$$\hat{y}^I = \text{softmax}(\boldsymbol{W_I}\boldsymbol{E^I} + \boldsymbol{b_I}) \tag{13}$$

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

In the above-mentioned formulas, $\boldsymbol{W_v}$, $\boldsymbol{W_{Mix}}$, $\boldsymbol{W_I}$ and $\boldsymbol{W_T}$ are trainable parameter matrices; $\boldsymbol{b_v}$, $\boldsymbol{b_{Mix}}$, $\boldsymbol{b_I}$ and $\boldsymbol{b_T}$ represent bias matrices.

3.5 Training

The training objective of our multi-modal metaphor detection model involves the integration of three distinct loss functions, denoted as $\mathcal{L}_I$, $\mathcal{L}_T$ and $\mathcal{L}_M$. The loss function is as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{D}_{\text{ME}}|} \sum_{i=1}^{|\mathcal{D}_{\text{ME}}|} L_{CE}\left(\hat{Y}, Y\right) \tag{15}$$

where $\mathcal{D}_{\text{ME}}$ is the number of samples in the dataset, The loss formula is parameterized as $\mathcal{L} = \{\mathcal{L}_I, \mathcal{L}_T, \mathcal{L}_M\}$, with $\hat{Y} = \{\hat{y}, \hat{y}^I, \hat{y}^T\}$ and Y representing the model's predicted outcomes and the true values, L_{CE} is the cross-entropy loss function.

To optimize the overall performance, we define the aggregate loss $\mathcal{L}_{sum}$ as a weighted combination of these individual losses. The final loss function is formulated as:

$$\mathcal{L}_{sum} = 0.5 \cdot \mathcal{L}_I + 0.5 \cdot \mathcal{L}_T + \mathcal{L}_M \tag{16}$$

4 Experiments

In this section, we begin by introducing the dataset used to validate our method, as well as the experimental setup. Following this, we report the experimental results and provide an analysis of these outcomes.

4.1 Data and Setting

We selected the multi-modal metaphor dataset proposed by Xu et al. (2022), which consists of 10,000 meme images collected from social media. Text information was extracted from these images using OCR methods to construct the multi-modal metaphor dataset, which includes 6,000 entries in Chinese and 4,000 in English. In addition to the classification labels for metaphors, they also annotated the source of the metaphors and their associated emotions.

All trained models were set with a learning rate of 1e-5, a batch size of 8, and were trained for 100 epochs with an early stopping mechanism in place. The dataset was randomly shuffled and divided into training, validation, and test sets in a 6:2:2 ratio. All experiments were conducted on a single 3090-24G GPU. The final results of our method were obtained by taking the average of five different random seeds, with the average single run time within 20-30 minutes. Finally, the model's performance was evaluated based on the F1 score.

The Low-Rank Adaptation (LoRA Hu et al. (2021)) fine-tuning approach was adopted for fine-tuning LLMs. All of the settings followed those used in Alpaca-LoRA*.

4.2 Baseline Methods

Language Models

We tested several common pre-trained models for this task, including the AutoEncoder M-BERT (Pires et al., 2019), XLM-R (Conneau et al.,

*Alpaca-LoRA

Modality	Model Type	Model	ACC	P.	R.	F1.
Language model	AutoEncoder	M-BERT-base	74.60	61.25	76.93	68.20
		XLM-R-base	83.32	78.57	72.71	75.53
	AutoRegressive Model	M-T5-base	83.86	80.25	71.91	75.85
		M-BART-large	83.52	78.79	73.14	75.86
	LLMs	LLaMA2-7b (LoRA)	83.07	78.23	72.29	75.15
		ChatGLM3-6b (LoRA)	84.81	82.22	72.86	77.26
Vision model	CNN Model	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
	Transformer Model	ViT-base	74.75	65.50	60.62	62.97
		Swin Transformer-base	78.83	77.82	56.26	65.31
Multi-modal model		VILT	83.13	78.01	72.86	75.35
		InternLM-XComposer-7b (Zero-shot)	67.50	30.83	17.29	22.16
		BLIP2-2.7b (Zero-shot)	38.33	33.44	82.97	47.05
		BLIP2-2.7b (LoRA)	85.66	80.61	78.34	79.46
Related Work		CLIP (Zhao et al., 2023)	75.05	60.83	**83.07**	70.23
		Vilio (Muennighoff, 2020)	84.30	79.97	79.97	76.74
		CoolNet (Xiao et al., 2023)	77.49	66.84	72.29	69.46
		MultiCMET (Zhang et al., 2023b)	85.66	82.69	75.25	78.79
		C4MMD (Ours)	**87.70**	**83.33**	81.58	**82.44**

Table 1: Results of different methods on the task of multi-modal metaphor detection.

2019), as well as the AutoRegressive models M-T5 (Xue et al., 2020) and M-BART (Liu et al., 2020). Additionally, we evaluated the capabilities of LLMs on this task by using LLaMA2 (Touvron et al., 2023) and ChatGLM3 (Zeng et al., 2022), due to their strong performance in both Chinese and English corpora. We fine-tuned both models separately using LoRA.

Vision Models

We also tested models from the vision domain, including Convolutional Neural Network (CNN) models such as VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and ConvNeXt (Liu et al., 2022), as well as models based on the Transformer architecture, like ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021).

Multi-modal Models

In the multi-modal model domain, we selected VILT (Kim et al., 2021), BLIP2 (Li et al., 2023a), and InternLM-XComposer (Zhang et al., 2023c) to test their capabilities in addressing the metaphor detection task. All three models employ the Transformer architecture, yet they differ significantly in model size. We tested the capabilities of these MLLMs both in a zero-shot setting and with LoRA fine-tuning.

Other Related Works

We also explored other works related to our task, thereby lending more credibility to our comparative analysis. Below, we introduce these works in detail.

- **CLIP** (Zhao et al., 2023): Evaluation of various models for hate meme detection task. We adopted best performance CLIP to evaluate its effectiveness in multi-modal metaphor detection tasks.
- **Vilio** (Muennighoff, 2020): An excellent method which achieves 2nd place in the Hateful Memes Challenge. It Uses OCR and entity recognition technologies to extract text and visual features from memes for better meme harmfulness detection tasks.
- **CoolNet** (Xiao et al., 2023): Extracting text syntactic structure to boost model's sentiment analysis ability on Twitter multi-modal data.
- **MultiCMET** (Zhang et al., 2023b): A baseline model for chinese multi-modal metaphor detection task. It uses the CLIP model to generate additional information to assist in the fusion between modalities.

4.3 Main Results

Table 1 shows the capabilities of different models in the task of multi-modal metaphor detection. Here we only evaluated the main classification results $\hat{y}$.

Model	ACC	P.	R.	F1.
Ours	87.70	83.33	81.58	82.44
-Fusion model	85.66	77.87	83.12	80.41
-CoT features	85.06	78.42	79.75	79.08
-Vision encoder	86.25	78.36	84.53	81.33

Table 2: Ablation study for the components in the model on metaphor detection.

VM	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

Table 3: The impact of different language and vision model combinations on the metaphor detection task, VM for Vision Model and LM for Language Model. We then use a linear layer to fuse the features of two modalities.

We did not assess the outcomes of the two subtasks $\hat{y}^I$ and $\hat{y}^T$ as the two subtasks which were primarily designed to serve the main task.

Our approach achieved the best results in both Chinese and English sample sets. Considering the outcomes produced directly by the MLLM (InternLM-XComposer-7b), we allowed it to indirectly generate additional features for images and texts, effectively leveraging the large model's capabilities. Coupled with a downstream classifier, this approach resulted in an additive effect.

The performance of multi-modal models varied widely, with most models not surpassing language models. This underscores the importance of textual modality in recognizing multi-modal metaphors. MLLMs did not perform well in zero-shot scenarios, partly due to our designed prompt templates. However, the primary reason is the models' inability to understand the task. Encouragingly, after fine-tuning BLIP2, its capabilities surpassed all other comparative methods. This demonstrates the benefit of interaction between image and text modalities in the task and how large models can effectively understand and address this task after fine-tuning.

In related work, studies closely aligned with our own, such as those by Zhang et al. (2023b) and Muennighoff (2020), have achieved competitive performances. However, Twitter sentiment classification by Xiao et al. (2023), which differs somewhat from our task, consequently showed weaker performance.

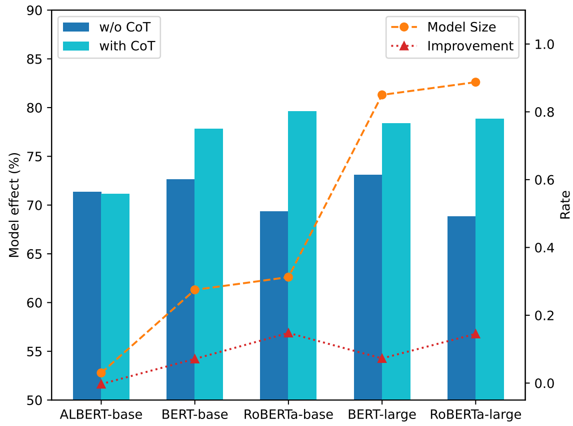

Figure 3: The effect of different sizes of models with or without CoT generation and the rate of improvement. We controlled the intercept of the model size between 0-1 to show the effect of improvement on a single figure.

4.4 Influence of Different Factors

Table 2 shows the effects demonstrated by our model after undergoing ablation experiments.

Replacing the fusion structure in the model with a linear layer resulted in a significant decrease in performance. This suggests the necessity of additional fusion structures to help the model understand the extra features generated by the MLLM. Moreover, eliminating the CoT generation method of the MLLM, and relying solely on a one-step generation method, led to an even more noticeable performance drop. This also indicates that the CoT method can generate better additional features, thereby assisting downstream models in making more accurate judgments.

Interestingly, the performance of the model declined only slightly when we removed the image processing module. This indicates that the MLLM can provide a certain level of visual information for smaller models, but more comprehensive information still requires the contribution of vision models.

4.5 The Impact of Different Language Vision Model Combinations

We tested the capabilities of multiple visual and textual models during modal fusion. The language model was uniformly set to MBERT when testing vision models and the ViT was used consistently when testing language models.

From the data in Table 3 and Table 1, although in single modality settings, the vision model VGG and the textual model M-T5 achieved the best per-

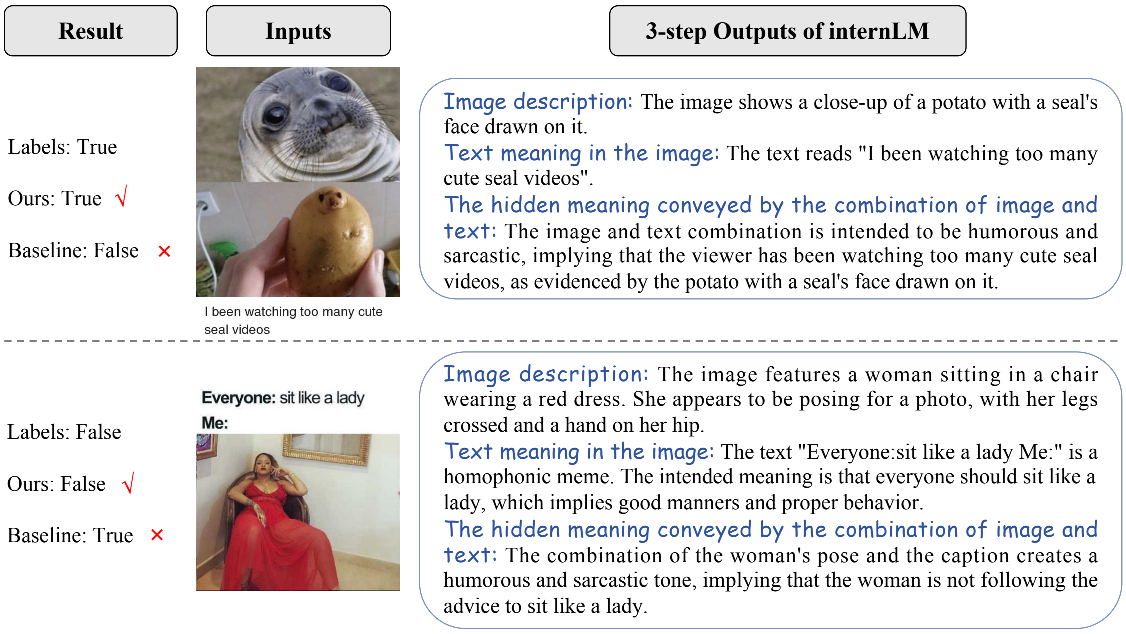

Figure 4: Examples of case study.

formance, the combination of ViT and XLM-R outperformed all others upon modal fusion.

The combinations of ResNet + MBERT and VGG + MBERT are also baseline models proposed by Met-Meme (Xu et al., 2022). According to the results, we reported the same results as them.

4.6 The Impact of Language Model Size

Figure 3 illustrates the abilities of models of different sizes under our architecture. Considering that the improvement ratio is generally between 0 and 1, while the model size is typically in the hundreds of millions, we divided all model sizes by 400 million to scale them between 0 and 1, allowing us to display both model size and improvement on the same graph. It was evident that as the model size increased, especially when the model was initially small, there was a progressively noticeable performance improvement. When the model was too small, the additional textual information did not yield positive effects; rather, it could have the potential to negatively impact the model's performance. It was only when the model size was increased that the model became capable of understanding longer contextual information.

4.7 Case Study

To further explore the effectiveness of our proposed model, we select two examples from the testing dataset illustrated in Figure 4.

The first example demonstrates an image-led metaphor. By directly comparing a seal with a potato, it depicts the consequences of looking at too many cute seals. The MLLM, through its understanding of the image, accurately recognized the resemblance between the seal and the potato, thereby aiding the downstream model in making the correct judgment.

In the second example, the MLLM identified features from both the image and text, and then combined these to correctly understand the humorous meaning expressed in the meme. The downstream model accurately recognized that it did not contain metaphorical features. In contrast, methods lacking the additional information from the large model judged it to be metaphorical based solely on the phrase "like a lady," leading to a misjudgment.

5 Conclusion

Our study aimed to tackle the challenges of multi-modal metaphor interpretation by leveraging advanced MLLMs. We designed a three-step method with CoT-prompting to extract richer information from both images and text. Augmented knowledge from MLLMs proved crucial in enhancing smaller models to grasp metaphorical features within each modality and in the fusion of modalities. This work not only advances multi-modal metaphor detection but also paves the way for future research exploring the potential of MLLMs in addressing complex language and vision challenges.

Limitations

We believe the main limitation of our work lies in only testing our metaphor detection ability within a multilingual meme dataset and not extending to other subtasks in meme datasets, such as harmfulness detection, nor to metaphor detection in other multi-modal datasets. However, despite the lack of experimental data, we are confident in our work's applicability in these directions, which will also be one of our future research focuses.

Additionally, regarding the meme dataset, we did not find a usage license, nor did we filter for potential harmfulness or offensiveness in the data, including in the extra features generated by the MLLM, which may contain toxic data, thus presenting a risk of offensiveness and harmfulness.

Although we used a method of averaging five tests for our model, for other comparative methods, we simply took the results from the first run for inclusion in our tables. We acknowledge this could introduce some error, but we believe that even if the comparative methods were tested in the same way, our method would still demonstrate overwhelmingly superior performance.

References

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang. 2022. Ring that bell: A corpus and method for multimodal metaphor detection in videos. *arXiv preprint arXiv:2301.01134*.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. *arXiv preprint arXiv:2305.17480*.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Findings of the association for computational linguistics: emnlp 2020*, pages 4479–4488.

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task. In *Proceedings of the second workshop on figurative language processing*, pages 235–243.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized 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. Unsupervised cross-lingual representation learning at scale. *arXiv preprint 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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational linguistics*, 17(1):49–90.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, pages 1–67.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian 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: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Gitit Kehat and James Pustejovsky. 2021. Neural metaphor detection with visibility embeddings. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 222–228.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. 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: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Shuqun Li, Liang Yang, Weidong He, Shiqi Zhang, Jingjie Zeng, and Hongfei Lin. 2021. Label-enhanced hierarchical contextualized representation for sequential metaphor identification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3543.

Yucheng Li, Shun Wang, Chenghua Lin, and Guerin Frank. 2023b. Metaphor detection via explicit basic meanings modelling. *arXiv preprint arXiv:2305.17268*.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. Cate: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. 2023. Metapro online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 127–135.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3888–3898.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PloS one*, 8(4):e62343.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 160–170.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the second workshop on figurative language processing*, pages 30–39.

Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng. 2023a. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7736–7752.

Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng. 2023b. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2887–2899.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Omnia Zayed, John P McCrae, and Paul Buitelaar. 2020. Contextual modulation for relation-level metaphor identification. *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. 2022. 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: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. 2023b. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association 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. 2023c. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. *arXiv preprint arXiv:2305.16638*.

Bryan Zhao, Andrew Zhang, Blake Watson, Gillian Kearney, and Isaac Dale. 2023. A review of vision-language models and their performance on the hateful memes challenge. *arXiv preprint arXiv:2305.06159*.