A Textual Modal Supplement Framework for Understanding Multi-Modal Figurative Language

Jiale Chen[1], Qihao Yang[1], Xuelian Dong[1], Xiaoling Mao[2] and **Tianyong Hao[1]***
[1] School of Computer Science, South China Normal University
[2] School of Languages and Literature, University of South China
{jlchen, charlesyeung}@m.scnu.edu.cn,
{xueliandong01, BriannaMxl}@163.com,
haoty@m.scnu.edu.cn

Abstract

Figurative language in media such as memes, art, or comics has gained dramatic interest recently. However, the challenge remains in accurately justifying and explaining whether an image caption complements or contradicts the image it accompanies. To tackle this problem, we design a modal-supplement framework MAPPER consisting of a describer and a thinker. The describer based on a frozen large vision model is designed to describe an image in detail to capture entailed semantic information. The thinker based on a finetuned large multi-modal model is designed to utilize description, claim and image to make prediction and explanation. Experiment results on a publicly available benchmark dataset from FigLang2024 Task 2 show that our method ranks at top 1 in overall evaluation, the performance exceeds the second place by 28.57%. This indicates that MAPPER is highly effective in understanding, judging and explaining of the figurative language. The source code is available at https://github.com/Libv-Team/figlang2024.

1 Introduction

Figurative language in media has gained much interests recently. By understanding similes and metaphors in the figurative language, it is possible to deepen the understanding of specific cultural contexts and social phenomena (Hwang and Shwartz, 2023). This task is challenging because it involves abstract reasoning about images, as well as it involves understanding social common sense and cultural contexts.

Traditional solutions extract features from images using CNNs and encode textual descriptions with RNNs (Mo et al., 2023; Chen et al., 2024), employ multi-modal fusion for inference (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), and determine and elucidate their interrelations through

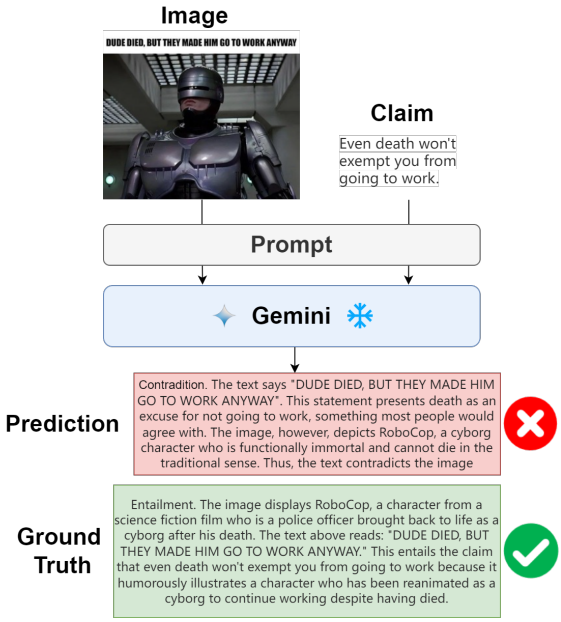

Figure 1: A typical method uses zero shot prompts to induce responses from a multi-modal large language model.

classification and explanation generation. With the development of Multi-modal Large Language Models (MLLM) in image captioning and Visual Question Answering (VQA), it turns to be a visual entailment task. The task first predicts whether an image caption entails the image or not and provide a text explanation for labeling prediction. New ideas also involve using formulated prompts according to heuristic rules to guide a large model in producing a relevant answer. The main framework of a typic method utilizing large model and prompt is shown in Figure 1.

Despite the progress made by these methods in dealing with visual entailment tasks, when faced with specific cultural and social contexts, the model ability to explain and reason is limited due to the lack of relevant context. Subsequently, the inconsistency between images and texts may make the models more challengeable to determine the entail-

*Corresponding author

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 85–91

ment relationships. Thereby, their performance is much worse than that of human beings.

To that end, we propose MAPPER (textual ModAl suPPlement framEwoRk), a figurative language understanding model. It consists of a describer and a thinker. The describer provides a textual model of the image as a modality supplement for further prediction and explanation by the thinker. Experiment result indicated that MAPPER is effective in understanding, judging and explaining of the figurative language. It shows that a simple fine-tuning method can significantly enhance the model performance in figurative language understanding with just minor prompt adjustments.

2 Related work

In recent years, advances in language modeling notably improved model comprehension of metaphorical language. Chakrabarty et al. proposed a model that fine-tuned T5 to understand metaphorical language through textual interpretation. Chakrabarty et al. introduced a knowledge augmentation model employing human strategies for explaining types of figurative language: inferring meaning from context and drawing on the literal meanings of constituent words. This knowledge augmentation model enhanced performance on discriminative and generative tasks, further narrowing the gap with human performance. Liu et al. created a Fig-QA benchmark through crowdsourcing for a broader study of metaphorical language. Their findings indicated that although pre-trained language models could achieve commendable performance after fine-tuning, their performance on a limited number of samples still fell significantly short of human capabilities.

In addition, with the development of multimedia, there had been an increased focus on generative understanding of multimodal metaphorical language. Hessel et al. investigated visual language models and language-only models for understanding multimodal metaphorical language and found that both types of models had difficulties in all three tasks. Desai et al. introduced an architecture based on a multimodal Transformer, which included a cross-modal attention mechanism focusing on the distinctive features between images and captions. This model obtained relatively high consistency scores in human evaluations. Yosef et al. utilized the state-of-the-art vision and language model CLIP (Radford et al., 2021) to perform on a multimodal metaphorical language comprehension task and found that it performed relatively poorly. The experimental results showed that the best model was only 22% accurate in the detection task, much lower than the 97% accuracy achieved by humans. This discrepancy was mainly due to the poor performance of model in understanding the connection between metaphorical language and images, with a tendency to prefer partially literal images over metaphorical ones.

These studies have primarily improved performance through methods such as model fine-tuning and knowledge enhancement. However, they still face challenges in understanding multimodal metaphorical language. To enhance the capability of visual language models to comprehend metaphorical language, we design prompts to clarify task requirements and employ modal supplement methods to boost the integration of multimodal data, aiming to narrow the gap between models and humans in multimodal metaphor comprehension.

3 The Method

The task of multimodal figurative language is defined as follows: Given an image claim C and an image I, a label L that indicate the caption entails or contracts to the image need to be predict. A corresponding explanation E of the predicted label is needed to be generated.

This paper proposes a textual modal supplement framework MAPPER, which is consisted of a describer and a thinker. The describer read the i-th image I_i, and used self-knowledge to describe the image as inherent thinking D_i. The thinker uses the inherent thinking D_i, image I_i and claim C_i to generate final predict L_i iand explanation E_i. The overview of the model is shown in Figure 2.

The Describer. To better understanding the image content, a MLLM-based describer is designed according the prompt instruction P_1 from the prompt template PTR_1 and the i-th image I_i to generate the image description D_i, It is worth noting that the parameter weights in MLLM are frozen. Formally:

$$D_i = MLLM_{frozen}\left(Prompt\left(I_i, C_i\right)\right) \quad (1)$$

Here, P_1 is designed as follows:

```
<Image> Please describe in detail what
you see in the provided image.
```

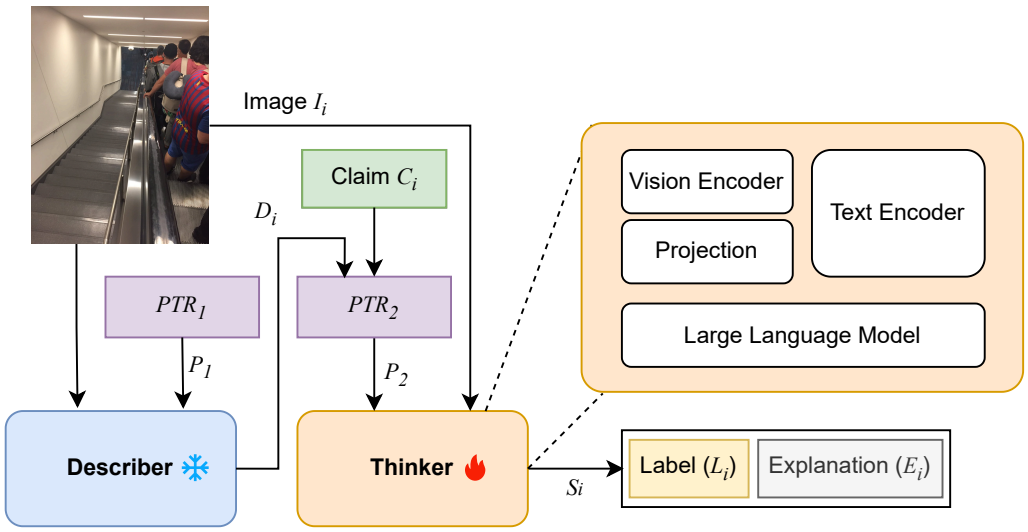

Figure 2: The overall architecture of our MAPPER framework with a describer and a thinker.

Data	Train/Valid		Test	
	Absolute	Proportion	Absolute	Proportion
Nycartoons	520	11.7%	87	12.6%
IRFL	1322	29.9%	198	28.7%
Muse	1000	22.6%	150	21.8%
Mamecap	853	19.3%	128	18.6%
Vismet	731	16.5%	126	18.3%

Table 1: Dataset statistics.

The Thinker. A prompt template PTR_2 is used firstly to generate a prompt based on the image description D_i and claim C_i. In this way, we unify the original classification and the generation tasks into one generation task. In this way, the thinker generates the responses S_i consisted of the concatenation of the label L_i and explanation E_i. The design of prompt template is shown as follows:

```
The description of this picture is
<Description>. The claim of this picture
is <Claim>. You need to predict the claim
of this picture is 'entailment' or
'contradiction' firstly according to the
picture and its description. Then you need
to give an explanation for the prediction.
The prediction and the explanation are
related to the meaning of the figurative
language expression. Your response must
follow the format shown as below:
"Prediction. Explanation".
```

Next, a vision encoder Enc_p is designed to encode the image I_i, and a text encoder Enc_e is used to encode the prompt p_2. $f(\bullet)$ is the projection function. The process is as shown in Equation 2–4.

$$H_v = f(Enc_p(I_i)) \tag{2}$$

$$H_l = Enc_e\left(Prompt\left(C_i, D_i\right)\right) \tag{3}$$

$$[L_i; E_i] = s_i = LLM\left(H_v, H_l\right) \tag{4}$$

Training. During training epoch, the model is trained as a minimized negative log likelihood as Equation 5.

$$\mathcal{L} = \sum_{j}^{n} -\log p\left(s_i^j \middle| s_i^{<j}, I_i, C_i, D_i\right) \tag{5}$$

s_i^j is the generated word output in the j-th time step that generated by the system. n is the maximum response length.

4 Experiments

4.1 Datasets

The V-FLUTE (Saakyan et al., 2024) was used in the experiments. It consisted of five small datasets, with data compiled from a series of prior work on visual metaphor and multimodal understanding, supplemented with annotated explanations detailing the implicit relationships (Yosef et al., 2023; Chakrabarty et al., 2023; Hwang and Shwartz, 2023; Hessel et al., 2023b; Jain et al., 2020; Shahaf et al., 2015). The statistical details of these datasets weres presented in Table 1. We followed datasets splits from the competition "UNDERSTANDING

Type	Method	Metrics		
		F1@0	F1@50	F1@60
Zero Shot	LLava-7B-v1.6 (offical baseline)	44.82	37.38	19.99
	LLava-7B-v1.5	43.40	40.42	20.30
	Gemini-Pro-Vision	59.57	58.61	42.36
	Gemini (Text only)	57.24	56.30	36.09
	GPT-4V	69.56	63.78	48.89
Fine-tune	TinyLLava-1.5B	72.56	71.39	59.24
	TinyLLava-3.1B	<u>86.12</u>	<u>85.40</u>	<u>71.56</u>
	MAPPER (Ours)	**89.67**	**89.09**	**74.15**

Table 2: Performance comparison of the models on the V- FLUTE datasets. The best performance is bolded and the second is underlined.

OF FIGURATIVE LANGUAGE THROUGH VISUAL ENTAILMENT" [1] for training, validation, and testing.

4.2 Evaluation Metrics

The evaluation metrics were primarily F1 scores for the label prediction. In addition, we used BERT-score (Yuan et al., 2021) to assess the quality of the explanation. Thus, the evaluation metrics were F1@0 (only F1 scores), F1@50 (computed F1 scores where only instances whose interpretations matched a reference with BERT-score higher than 50 were treated as correct), and F1@60 (computed F1 scores where only instances whose interpretations matched a reference with BERT-score higher than 60 were treated as correct). These metrics were based on previous work in FigLang2022.

4.3 Baseline

Three categories of baseline models were evaluation in this experiment. 1) The origin multi-modals models: Gemini-Pro-Vision (Team et al., 2023), and GPT-4V. 2) The models consisted of an image encoder and a large language model: LLava-7B (Liu et al., 2023), TinyLLava (Zhou et al., 2024). 3) The large language model: Gemini-Pro.

4.4 Hyperparameters

A frozen parameter LLava-7B-v1.5 model was used for the describer, while a finetuned LLava-7B-v1.5 model with Lora (Hu et al., 2021) for the thinker. The training epoch was set to 3. The batch size was set to 4 and the learning rate was set to $2.5e^{-5}$. The rank of the Lora model was set to 128. The learning rate scheduler type was used "cosine", and the max length of model was constraint to 2048. The vision tower of MAPPER used CLIP. The warm up ration was set to 0.03. All experiments were conducted in a NVIDIA 4090 GPU with 24GB memory.

5 Results and Analysis

5.1 Main Result

The results of comparing with the baseline models were shown in Table 2. It could be seen that our MAPPER achieved the highest scores on all three metrics through the supervised fine-tuning method. Specifically, F1@0 reached 90.06, F1@50 reached 89.49, and F1@60 reached 76.33. Compared with the Prompt-based Zero Shot method, our method improved the performance of these three indicators by 28.57%, 34.33%, and 50% respectively after supervised fine-tuning compared with the method without fine-tuning. These high F1 scores indicated that MAPPER could effectively understand the information in images and give accurate predictions and explanations. Moreover, some interesting phenomena were identified. The Gemini-Pro-Vision model performed better than the Gemini (Text only) model in zero-shot conditions. This might indicate that replacing images with their descriptions could cause information losing in predication, resulting in worse performance.

5.2 Ablation Study

To explore the effectiveness of each component, an ablation study was conducted. We removed describer and PTR_2 of the MAPPER one by one and then analyzed the performance changes. The result was shown in Table 3. From the table result,

[1] https://www.codabench.org/competitions/1970/#/pages-tab

Method	F1@0	F1@50	F1@60
Ours	**89.67**	**89.09**	**74.15**
- Describer	89.20 (0.47 ↓)	89.04 (0.05↓)	75.58 (1.43 ↑)
- PTR_2	51.16 (38.51 ↓)	50.46 (38.63 ↓)	41.45 (32.70 ↓)

Table 3: The ablation experiment with/without two main modules Describer and PTR_2.

Image:  Claim: The cat is busy as a bee

Method	Prediction + Explanation	F1	BERT-score
MAPPER	Entailment. The image depicts a cat lying on a bed with glasses on its face, appearing to use a laptop with its paws, as if it were working or using the computer. The simile "The cat is as busy as a bee" is entailed by the image because the cat is engaged in an activity that resembles work or productivity, similar to how a bee would be busy with its tasks.	1	81.85
TinnyLLava-3.1B	Entailment. The image shows a cat sitting on a bed with a laptop, wearing glasses and a bowtie, as if it were a human working on the computer. The simile "The cat is as busy as a bee" is entailed by the image because the cat's posture and attire suggest that it is engaged in a task, similar to how a bee is industrious and focused on its work.	1	80.98
TinyLLava-1.5B	Entailment. The image depicts a cat wearing glasses and a tie, sitting on a bed with a laptop in front of it. The simile ""The cat is as busy as a bee"" is entailed by the image because the cat's attire and the context suggest that it is engaged in a task that requires focus and attention, similar to the industrious nature of a bee."	1	79.44
Ground Truth	Entailment. The image shows a cat with glasses on, looking at a laptop as if it's working or engaged in an activity. The simile "The cat is as busy as a bee" is entailed by the image because the cat appears to be actively involved with the laptop, mimicking the busy behavior of a bee with human-like working characteristics.	-	-

Figure 3: MAPPER and two baseline generated explanations in the V-FLUTE dataset.

we could draw the following observations:

The describer had relatively little impact on model performance. Without "Describer", the model performance scores on F1@0 and F1@50 dropped by 0.47 and 0.05 respectively, which were a relatively small change. However, it was worth noting that the F1@60 score increased by 1.43, which might indicate that the describer might had limitations when dealing with complex or difficult-to-classify cases.

The PTR_2 had a large impact on model performance. When we removed PTR_2, the model performance scores on all three metrics dropped significantly, especially on F1@0 and F1@50, where the scores dropped by 38.51 and 38.63 respectively. This indidated that the PTR_2 component played a key role in the model and had a significant impact on the model performance.

Overall, these results indicated that the performance of our MAPPER relied heavily on the PTR_2, while the describer component had a relatively less impact. This provided us with important guidance when improving the model and optimizing performance in the future.

5.3 Case Study

Figure 3 presents a case study demonstrating the comparesion between MAPPER and two baselines. The data of this case study was sourced from the V-FLUTE dataset. The label predictions and explanations were generated by MAPPER and two

baseline methods, TinyLLava-1.5B and TinyLLava-3.1B. The input consisted of an image and a claim. The image showed a cat working in front of a computer accompanied by the claim "That cat is busy as a bee". All three methods accurately predicted the labels. However, the explanation generated by MAPPER achieved the highest BERT-score compared to the baseline methods. There were some biased words using in the baseline explanations in contrast to the ground truth. Our MAPPER explicitly indicated the "busy" and generated explanation more closely resembling the ground truth, resulting in the highest BERT-score. This case exemplifies capacity of MAPPER to generate explanations that closely align with the ground truth.

6 Conclusion

This paper proposed a textual modal supplement method MAPPER for figurative language understanding. The MAPPER used a frozen LLava as the describer to generate a description of the image and a finetuned MLLM as the thinker to make predictions and explanations for the figurative language within image and claim. Experiment results on the public datasets indicated that our MAPPER achieved the state-of-the-art performance. The results illustrated that a finetune in small dataset about understanding of figurative language could highly improve MLLM model performance.

Limitation

Due to competition time constraints, we did not explore clearly in this experiment why the textual modal supplement generated by describer can have a negative impact on F1@60. Although our method ranked first in the competition, this paper did not design different prompts to test the robustness of our method. In addition, we did not further explore whether a MLLM with a larger number of parameters can learn more accurate judgment and understanding of figurative language in pictures.

Acknowledgments

The work is supported by grants from National Natural Science Foundation of China (No. 62372189) and the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E09/22).

References

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Guanhua Chen, Qiqi Xu, Choujun Zhan, Fu Lee Wang, Kai Liu, Hai Liu, and Tianyong Hao. 2024. Improving open intent detection via triplet-contrastive learning and adaptive boundary. *IEEE Transactions on Consumer Electronics*.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023a. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023b. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the ACL*.

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*.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert. 2020. The New Yorker cartoon caption contest dataset.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Di Mo, Bangrui Huang, Haitao Wang, Xinyu Cao, Keqin Gan, Jie Wei, Heng Weng, and Tianyong Hao. 2023. Sclert: A span-based joint model for measurable quantitative information extraction from chinese texts. *IEEE Transactions on Consumer Electronics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-FLUTE: Visual Figurative Language Understanding with Textual Explanations Dataset. https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE. Dataset associated with the paper "V-FLUTE: Visual Figurative Language Understanding with Textual Explanations".

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *KDD*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.