LMEME at SemEval-2024 Task 4: Teacher Student Fusion - Integrating CLIP with LLMs for Enhanced Persuasion Detection

Shiyi Li*, Yike Wang*, Liang Yang†, Shaowu Zhang, Hongfei Lin
Dalian University of Technology
{lishiyieee, yike}@mail.dlut.edu.cn,{liang, zhangsw, hflin}@dlut.edu.cn

Abstract

This paper describes our system used in the SemEval-2024 Task 4 Multilingual Detection of Persuasion Techniques in Memes. Our team proposes a detection system that employs a Teacher Student Fusion framework. Initially, a Large Language Model serves as the teacher, engaging in abductive reasoning on multimodal inputs to generate background knowledge on persuasion techniques, assisting in the training of a smaller downstream model. The student model adopts CLIP as an encoder for text and image features, and we incorporate an attention mechanism for modality alignment. Ultimately, our proposed system achieves a Macro-F1 score of 0.8103, ranking 1st out of 20 on the leaderboard of Subtask 2b in English. In Bulgarian, Macedonian and Arabic, our detection capabilities are ranked 1/15, 3/15 and 14/15.

1 Introduction

Memes are one of the most popular content types in online disinformation activities. They thrive on social media platforms, effortlessly reaching vast audiences. Memes within disinformation activities employ various rhetorical and psychological techniques, such as oversimplification of causation, insults, and defamation, to achieve their impact on users. In this context, meme detection is crucial for identifying and reducing the spread of false information.

The SemEval-2024 Task 4 (Dimitrov et al., 2024) is a multilingual detection task involving persuasion techniques in memes, and we participate in Subtask 2b, the binary classification task, to identify whether it contains a persuasion technique or no technique.In addition to English, the task also includes three test datasets in different languages, which are only released together with the test data in the final phase of the task.The purpose of this design is to evaluate the model's performance in zero-shot scenarios, specifically its capability on languages it has not encountered before.

The key to meme detection lies in uncovering rich correlations within memes between seemingly unrelated text and image components, particularly when there is no apparent connection between the text and image. In cases where the implicit meaning needs deeper exploration and understanding, traditional detection methods often fall short, as they approach meme detection in a straightforward end-to-end manner, overlooking a profound comprehension of meme text and images. Recently, Large Language Models (LLMs) have found success in complex reasoning. They could reveal the underlying implicit meanings beneath the surface of memes, enabling the assessment of whether persuasion techniques are present. Inspired by heuristic teaching (Pintrich and Schunk, 1996), where a teacher with rich experience can impart to students correct thinking and reasoning based on questions and corresponding answers, the students then learn how to deduce their own ways to the correct answers from questions.

To better harness the powerful reasoning capabilities and knowledge reservoir of LLMs, we proposed a Teacher Student Fusion detection system based on the CLIP (Radford et al., 2021) model and LLMs. This system operates in two stages: in the first stage, as the teacher model, the LLM is used to extract prior background knowledge related to persuasion techniques from memes; in the second stage, leveraging this prior knowledge, we fine-tune a smaller student model to detect whether memes contain persuasion techniques.

2 Related Work

2.1 Meme Classification Methods

Meme classification has emerged as a rising multimodal task in recent years. Early multimodal ap-

*These authors contributed equally to this work.
†Corresponding author.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 628–633

proaches include models like concatBERT (Kiela et al., 2019), which simply concatenates features from both images and text. Li et al. (2019) introduce the VisualBERT model, touted as the first image-text pretraining model. It utilizes Faster RCNN (Girshick, 2015) for image feature extraction, combines the extracted image features with text embeddings, and then inputs the concatenated features into a single Transformer structure initialized by BERT (Devlin et al., 2018) for classification. Lee et al. (2021) propose the DisMultiHate model, which enhances the classification capability and interpretability of hate memes by introducing entity detection in memes and incorporating statistics on race and gender information as supplementary data. Zia et al. (2021) employ the CLIP encoder to obtain features from both images and text, then simply concatenate these features and pass them to a logistic regression classifier. However, current solutions only capture the superficial signals of different modalities in memes in an end-to-end manner, failing to guide the model in-depth understanding of the complex and diverse relationships between visual and textual elements.

2.2 Large Language Models

Recently, LLMs (Brown et al., 2020) have demonstrated remarkable reasoning capabilities, generating high-quality reasoning steps to augment input prompts to LLMs and improve their few-shot or zero-shot performance (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022b). Reasoning steps have also been employed for additional fine-tuning to "self-improve" LLMs (Zelikman et al., 2022; Huang et al., 2022). Unfortunately, the large size of LLMs restricts their deployment on detecting memes with diverse modalities. Knowledge distillation has been successfully used to transfer knowledge from larger, more competent teacher models into smaller student models affordable for practical applications (Buciluǎ et al., 2006; Hinton et al., 2015; Beyer et al., 2022). However, existing researches on knowledge distillation from LLMs (Wang et al., 2022a; Ho et al., 2022; Magister et al., 2022) only consider the language modality. To accommodate multimodal features, we conduct abductive reasoning from LLMs, extracting underlying rationales as prompt arguments to assist in meme detection when fine-tuning smaller language models (LMs) for meme detection.

3 System overview

We define a meme detection dataset that potentially contains persuasion techniques as a set of memes where each meme $M = \{I, T, y\}$ is a triplet representing a visual content I that is associated with the textual T, and a ground-truth label $y \in \{propagandistic, non\_propagandistic\}$.

The core idea of our teacher student fusion model is to reason and develop a cognition-level rationale beyond the recognition-level perception (Davis and Marcus, 2015) by constraining the relationships between visual and textual elements in memes. To better utilize multimodal reasoning distilled from LLMs, this task is formulated as a natural language generation paradigm, where our model takes the text T and image I as the input and generates a textual output of the label y to clearly express whether at least one persuasion technique is present in the meme or not. In this paper, we propose to utilize abductive reasoning from LLMs with multimodal inputs to train smaller downstream models. Our overall framework is illustrated in Figure 1, which consists of abductive reasoning from LLMs and model fine-tuning.

3.1 Stage 1: Abductive Reasoning from the Teacher Model

We activate explicit reasoning knowledge in LLMs as the teacher model. Through prompt learning in causal reasoning, the LLM acquires meme-related context and hidden information, to guide student model in detecting persuasion techniques. Given a meme sample M from the training data, we first extract the text caption $\hat{I}$ of the image I to represent the visual content by off-the-shelf captioning model[1]. Specifically, based on the triplet $\left\{\hat{I}, T, y\right\}$, we design a prompt p:

"*Given the textual of a meme: [T], which is embedded in its image: [$\hat{I}$], labeled as [y]. Please provide a streamlined rationale for inferring the meme as [y], incorporating prior background knowledge related to persuasion techniques but without explicitly indicating the label.*"

It prompts the LLMs to generate a rationale R including rich contextual background knowledge, enabling the inference of whether persuasion techniques are present in memes.

[1] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

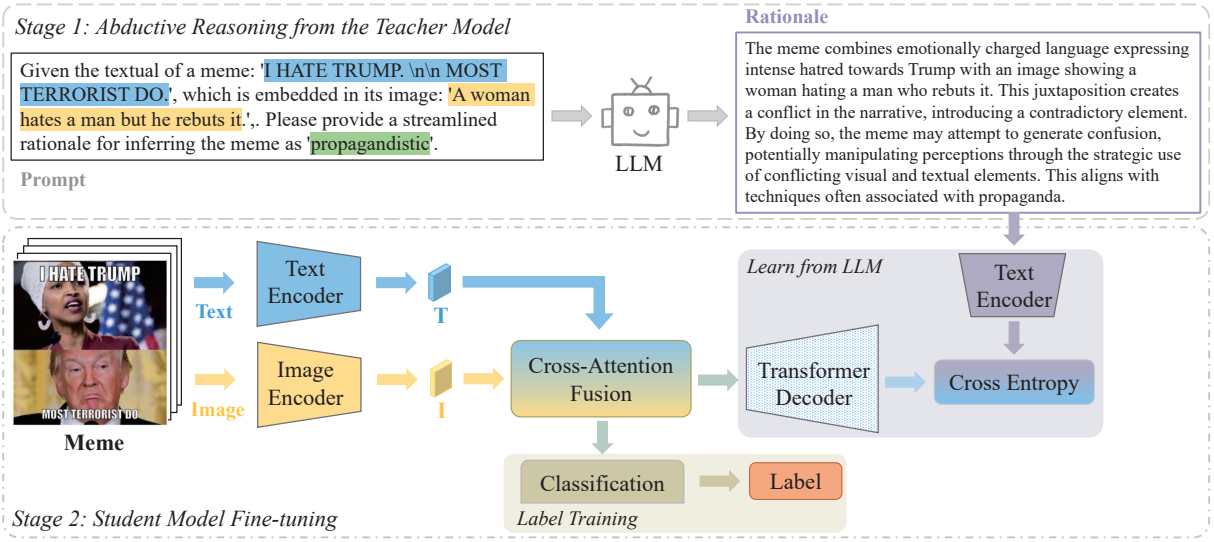

Figure 1: The framework of our method involves two stages. We first conduct abductive reasoning from the teacher model to extract rationales (purple) by the prompt consisting of the meme text (blue), the image caption (yellow), and the label (green). Utilizing the generated rationales, we then train the student model and predict whether memes contain persuasion techniques.

3.2 Stage 2: Student Model Fine-tuning

To facilitate the interaction between meme text and images, we fine-tune a smaller student model for persuasion detection tasks. By using the reasons generated by the teacher model as background knowledge, we aid in uncovering the rich interrelationship between text and vision modalities of memes. For a meme sample M,we first use the encoder of the CLIP model to encode the text T and image I input to obtain the embedding vector H_T and H_I.The advantage of attention mechanism in modality fusion and alignment lies in its ability to dynamically allocate and adjust weights for different modalities, allowing the model to flexibly focus on specific parts of the input. By emphasizing relationships between modalities, the attention mechanism helps improve the effectiveness of modality fusion and enhances the model's ability to capture important information. Therefore, we adopted a cross-attention mechanism for the fusion of textual and visual features.

$$Q_T = W_Q H_T + b_Q \quad (1)$$

$$K_I = W_K\ H_I + b_K \quad (2)$$

$$V_I = W_V\ H_I + b_V \quad (3)$$

$$H_o = Softmax\left(\frac{(K_I)^T Q_T}{\sqrt{d}}\right) V_I \quad (4)$$

Among them, d is the dimension of the feature, softmax is the activation function.

Learn from LLM By inputting the fused features H_o to the transformer decoder for decoding,we obtain the decoded output. Subsequently, we calculate the cross-entropy loss between this output and the given reason, as expressed by the following formula. Minimizing the cross-entropy loss between the meme feature and the generated reason feature facilitates the extraction of prior knowledge from the generated reason. This process helps the model transfer the contextual background information from the generated reason to the meme feature.

$$\mathcal{L}_{llm} = CrossEntropy(decoder(H_o), R) \quad (5)$$

Label Training Through the aforementioned process, our model has acquired the reasoning ability to extract persuasion techniques from LLM. As the objective of the task is to determine whether a meme contains persuasion techniques, label prediction becomes essential. This process shares the same model architecture as the preceding steps, with the introduction of a classifier during the decoding phase for label prediction. During the training process, we optimize the model by minimizing the cross-entropy loss between the predicted labels and the ground truth labels. The loss function is expressed as follows.

$$\mathcal{L}_{label} = CrossEntropy(y^{pre}, y^{true}) \quad (6)$$

Through the above process, for the data samples to be predicted, we can directly predict the labels

Parameter	From LLM	Label Training
Epochs	20	10
Batch size	32	32
Learning rate	5e-4	5e-5
Warmup step	0.1	0.1
Warmup Strategy	Linear	Linear
Image size	224	224

Table 1: Hyper-parameters.

Dataset	Model	Macro-F1
English - Dev	baseline	0.2500
	LMEME(w/o llm)	0.8329
	LMEME	0.8428
English - Test	baseline	0.2500
	LMEME(w/o llm)	0.8043
	LMEME	0.8103

Table 2: Main experimental results of Subtask 2b in English. LMEME is the model proposed in our study. LMEME(w/o llm) represents the ablation results without rationales generated by the teacher model. The baseline refers to the evaluation's benchmark.

of the model without generating corresponding rationales.

4 Experiments

4.1 Dataset and Evaluation

The dataset from the Task 4 of SemEval-2024 contains memes potentially employing persuasion techniques. After training on the English dataset, the model is tested across four languages: English, Bulgarian, North Macedonian and Arabic. As mentioned in the official task description, we employ macro-F1 to evaluate the performance of binary classification Subtask 2b.

In the experiments, we consider the 7b LLaMa2 model (Touvron et al., 2023) as the teacher model. For the task-specific student model, we utilize the CLIP-ViT-B/32 (Radford et al., 2021) architecture as the foundational framework. During the training phase, we evaluate the performance of the model every 100 steps and retain the parameters of the model that performed best on the validation set. The hyperparameters settings adopted are detailed in Table 1. All models are trained on NVIDIA GeForce GTX 3090 GPU.

Dataset	Model	Macro-F1
Bg - Test	baseline	0.1667
	LMEME(w/o llm)	0.6250
	LMEME	0.6710
NM - Test	baseline	0.0909
	LMEME(w/o llm)	0.5536
	LMEME	0.5908
Ar - Test	baseline	0.2271
	LMEME(w/o llm)	0.2933
	LMEME	0.3620

Table 3: Main experimental results in Bulgarian, North Macedonian and Arabic.

4.2 Results and Analysis

Table 2 shows the English detection capabilities of our system in Persuasion Techniques in Memes. A noticeable improvement is observed when comparing it to the scenario without the incorporation of prior knowledge from the LLM, our system in this study demonstrates superior performance. This indicates that leveraging background knowledge obtained from the teacher model can enhance the model's understanding of persuasion techniques to some extent, assisting the model in more accurately detecting memes. Furthermore, the combination of CLIP's encoding capability and the design of cross attention fusion using attention mechanisms enables the system to better align semantic features between text and visuals, facilitating more effective meme persuasion detection.

Table 3 displays our system's zero-shot capability in Bulgarian, North Macedonian, and Arabic. It can be observed that our system exhibits a significant improvement over baseline results in Bulgarian and North Macedonian, ranking 1/15 and 3/15 in the task. In comparison, there is also an improvement in results for Arabic, although not as pronounced as in the first two languages. The potential reason for this lies in the fact that English is an inflectional language with some agglutinative and analytic features. Bulgarian and Macedonian also share some features as inflected languages. However, Arabic introduces agglutinative features onto its inflectional foundation. Since our model has been trained on an English dataset, its effectiveness in detecting agglutinative features might be slightly inferior compared to inflectional languages. Additionally, the ablation results on these

three datasets also indicate the superiority of introducing prior knowledge. This further validates the effectiveness of the teacher student fusion system proposed in this paper.

5 Conclusion

This paper provides a detailed exposition of our approach in addressing Subtask 2b of Semeval2024 Task 4. Our teacher student fusion system initially leverages the Large Language Model as the teacher model to generate background knowledge regarding whether memes contain persuasion techniques. Subsequently, we incorporate this knowledge to fine-tune the student model by minimizing cross-entropy loss, sharing learned parameters with the predictive parameters of the model. Finally, we proceed with predictions using the trained model.

In the future, we will explore alternative encoding methods, better aligned image and textual semantic fusion methods, LLMs of different sizes and types, and different ways of prompting LLMs to generate enhanced background knowledge for meme persuasion detection.

References

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Paul R Pintrich and Dale H Schunk. 1996. Motivation in education: Theory, research, and applications.

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.

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*.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. *arXiv preprint arXiv:2211.01562*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219.