Multimodal Chain-of-Thought Reasoning for Social Fake News Detection

Xingchen Ding, Chong Teng, Donghong Ji
School of Cyber Science and Engineering,
Wuhan University, Wuhan, China
Email: {xingos,tengchong,dhji}@whu.edu.cn

***Abstract*—Social media has both positive and negative effects on news consumption. While it offers low cost, easy access, and rapid dissemination of information, it also facilitates the spread of fake news. Therefore, detecting fake news is a crucial challenge. In recent times, researchers in social forensics have shown significant interest in multimodal fake news detection (MFD). Although several approaches have incorporated customized attention mechanisms to facilitate the fusion of unimodal features, there are still unresolved issues regarding the optimal integration of multimodal features and its impact on decision-making in MFD. Furthermore, the potential benefits of pretrained language models (PLMs) for MFD remain largely untapped. In this paper, we propose a novel approach called Multimodal Chain-of-Thought Reasoning for MFD (MFDCoT) that explicitly understands the content of news with images to enhance the performance of MFD models. Specifically, we use multimodal chain-of-thought reasoning to comprehend news content on PLMs and employ an improved contrastive prompt learning paradigm in the MFD task. Extensive experiments have been conducted on two multimodal fake news datasets sourced from real-world environments. The results of our experiments show that our proposed framework outperforms state-of-the-art methods and achieves superior detection performance, even in the zero-shot setting.**

***Index Terms*—Few-Shot, Fake news detection, Social media, Multimodal, Chain-of-Thought, Prompt learning**

I. Introduction

The detection of fake news is crucial to prevent its rampant dissemination on social media and the Internet. Generally, the task involves identifying fake news based on extracted features from various sources, such as textual contents, attached images, and social contexts. Early works on fake news detection focused solely on analyzing text-only or image-only content [1], [2], [3], [4]. These works often employ a pre-trained model to verify the logical and semantic soundness of the input. While single-modal fake news detection schemes are effective, they do not account for the correlation between multiple modalities present in modern news and posts.

In recent years, there have been many works that incorporate multimodal features to detect anomalies in news and posts [5], [6], [7]. For example, Spotfake+ [5] integrates pre-trained language transformers and ImageNet models using multiple layers, while SAFE [6] jointly learns representations of news and visual information along with their relationship. Chen [7] additionally trains variational autoencoders (VAE) that learns to minimize the Kullback-Leibler (KL) divergence for news with correct image-text pairs. However, a drawback

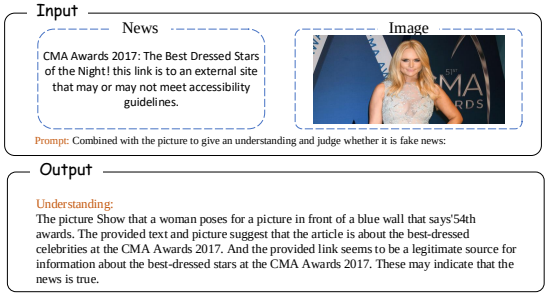

Fig. 1. Example of the understanding generation of news with images.

of these multimodal methods is that they overlook cross-modal correlation knowledge, which may lead to suboptimal results. As mentioned in [6] and [7], the correlations between representations from different modalities are crucial in MFD. Moreover, the corresponding signal spaces of textual and visual networks are different, which may negatively impact performance when directly learning a shared feature between the textual and visual networks.

Building on the success of Prompt Tuning (PT) [8], which effectively leverages pretraining language models by maintaining the same learning objectives as pretraining, and Chain-of-Thought (CoT) [9], which improves the ability of large language models to perform complex reasoning via intermediate reasoning steps, we propose a novel approach for MFD, called Multimodal Chain-of-Thought Reasoning (MFDCoT). In particular, we generate an understanding of news with images by inputting news text and news-related images into a pretraining language model (PLM), as illustrated in Figure 1. To better leverage the generated news understanding based on multimodal information, we propose a novel supervised contrastive prompt tuning paradigm. Our extensive experiments on two real-world multimodal fake news datasets confirm that our proposed model achieves outstanding performance for detecting multimodal fake news in few-shot settings compared to state-of-the-art baselines with a large margin. Furthermore, our method performs particularly well in zero-shot fake news detection.

II. Related Work

Multimodal Fake News Detection. Numerous unimodal research works have been proposed for detecting fake news [1], [2], [3], [4]. Although these unimodal characteristics

play a vital role in distinguishing fake news, the multimodal characteristics such as correlation and consistency are often ignored, potentially impairing the overall performance of these unimodal schemes on multimodal news.Previous literature has focused on mining useful representations from images and texts of the news for fake news detection. Earlier works have designed sophisticated yet black-box attention mechanisms for multimodal feature fusion [10], [11]. Other works have proposed to better align the extracted features from different modalities before sending them into the classifier [5], [6], [7], [12]. Dhruv [13] processed the image and text using unimodal feature extractors and further utilized a multimodal VAE to learn a shared representation from them. The sampled representation produced by the VAE is then sent to a decoder, which attempts to reconstruct the original texts and low-level image features. Zhou [12] used CLIP-based learning and a modality-wise attention mechanism to measure the cross-modal similarity and guide the mapping and fusion of features.

Despite achieving decent performance in MFD, there are still issues to be addressed. Firstly, understanding the content of both images and texts within news remains unclear. Secondly, little work in fake news detection considers applying the recently emerged arts in multimodal learning, motivating the use of multimodal chain-of-thought and contrastive prompt tuning to further boost performance.

CoT Reasoning with PLMs. Recently, the use of CoT has gained widespread popularity for assessing the multi-step reasoning capabilities of PLMs [9]. Specifically, CoT techniques facilitate the generation of intermediate reasoning chains by PLMs to solve a given problem. Empirical evidence suggests that PLMs are capable of performing CoT reasoning through Few-Shot-CoT paradigms [9], [14]. In Few-Shot-CoT, a small number of step-by-step reasoning demonstrations are employed as inference conditions. Each demonstration comprises a question and a reasoning chain that leads to the final answer, which is obtained via manual crafting or automatic generation. These techniques are respectively known as Manual-CoT [9] and Auto-CoT [14]. However, existing research on CoT reasoning primarily focuses on text language modality and gives little attention to multimodal scenarios.

Prompting for PLMs. Prompt-based learning has gained immense popularity in recent times for extracting knowledge from large language models, as evident from several recent research studies [15], [16] that have focused on prompt-tuning. Handcrafted prompts have traditionally been used to achieve impressive performance in prompt-based learning, particularly with the advent of GPT-3. More recent approaches such as AutoPrompt [17] and LM-BFF [18] have proposed automatic prompt construction through the generation of discrete tokens. Unlike previous works on unimodal fake news detection that were prompt-based [19], [4], our framework places a primary emphasis on detecting multimodal fake news. Specifically, our approach leverages multimodal chain-of-thought reasoning to understand news articles that contain images. This innovative approach represents a novel exploration of the challenging task of detecting multimodal fake news.

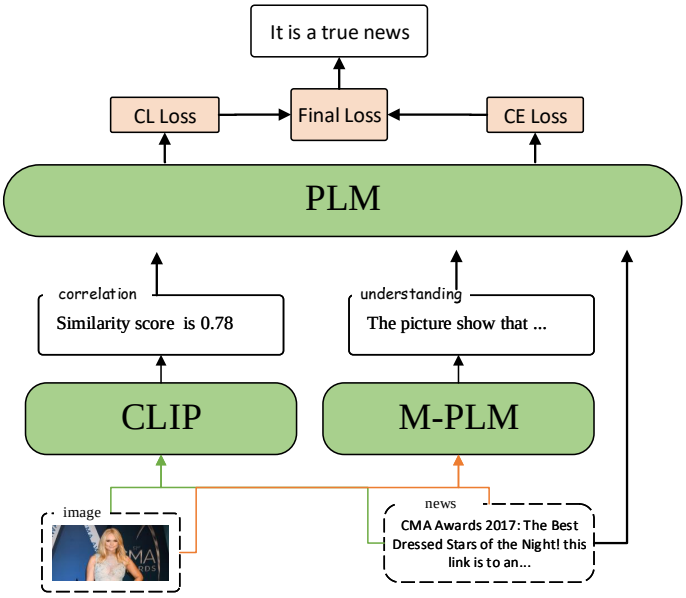

Fig. 2. A high-level illustration of the proposed MFDCoT framework for MFD.

Contrastive Learning. Contrastive learning (CL) aims to enhance the process of representation learning by increasing the agreement among instances of the same type and distinguishing them from others of different types [20]. In recent years, CL has demonstrated significant success in unsupervised visual representation learning [21], [20]. In addition to computer vision, recent studies [18], [22], [23] suggest that CL holds promise in various areas such as semantic textual similarity, stance detection, and short text clustering. However, the existing CL frameworks are primarily designed to augment unstructured textual data, such as sentences and documents, which makes them unsuitable for MFD tasks that involve news containing both textual and visual data.

III. Proposed Method

In this section, we present the proposed MFDCoT framework, as depicted in Figure 2, which aims to comprehend news articles that contain images, by leveraging a multimodal chain-of-thought reasoning technique trained on news. The proposed method significantly improves the identification of fake news.

A. Understand News with Images

Correlation understanding. CLIP, a multimodal model that combines knowledge of language concepts with semantic knowledge of images, is capable of predicting the most relevant text snippet given an image and vice versa. Together with other advanced multimodal technologies, CLIP can prove to be beneficial in fusing image-text features. To address the correlation issue between the multimodal features, we measure the cosine similarity between the text features and the image features provided by CLIP. The cosine similarity is calculated using the following equation:

$$\text{sim} = \frac{p_{news} \cdot (p_{image})^T}{|p_{news}|\,|p_{image}|}, \tag{1}$$

After calculating the cosine similarity, we then apply standardization and a sigmoid function to map the similarity into the range of 0 to 1. To perform the normalization, we calculate the running mean and standard deviation during training, and then subtract the running mean from the similarity score and divide it by the running standard deviation.

Content understanding. The probability of generating news-understanding text Y of length N, given news with prompt input X_{news} and image input X_{image}, is computed using Equation 2. We fine-tunes a Transformer-based FlAN-T5 (M-PLM) [24] to implement $p_\theta\left(Y_i \mid X_{\text{news}}, X_{\text{image}}, Y_{<i}\right)$.

$$p(Y|X_{\text{news}}, X_{\text{image}}) = \prod_{i=1}^{N} p_\theta\left(Y_i \mid X_{\text{news}}, X_{\text{image}}, Y_{<i}\right), \quad (2)$$

The M-PLM $F(X)$ uses both news and image inputs to obtain the text representation H_{news} and the image feature H_{image}, as shown in Equation 3 and Equation 4. The NewsEncoder$(\cdot)$ function is implemented using a Transformer model, while the ImageExtractor$(\cdot)$ function vectorizes the input image into vision features through DETR [25].

$$H_{\text{news}} = \text{NewsEncoder}(X_{\text{news}}), \quad (3)$$

$$H_{\text{image}} = W_h \cdot \text{ImageExtractor}(X_{\text{image}}), \quad (4)$$

Subsequently, we employ a single-head attention network to correlate text tokens with image patches. The query ($\mathcal{Q}$), key ($\mathcal{K}$), and value ($\mathcal{V}$) are H_{news}, H_{image}, and H_{image}, respectively. The attention output $H_{\text{image}}^{\text{attn}} \in \mathbb{R}^{n\times d}$ is computed using Equation 5, where d_k has the same dimension as H_{news} since we use a single head.

$$H_{\text{image}}^{\text{attn}} = \text{Softmax}(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d_k}})\mathcal{V}, \quad (5)$$

After obtaining news and image representations, we use the gated fusion mechanism to fuse H_{news} and H_{image}. The fused output $H_{\text{fuse}} \in \mathbb{R}^{n\times d}$ is calculated using Equations 6 and 7, where W_l and W_v are learnable parameters.

$$\lambda = \text{Sigmoid}(W_l H_{\text{news}} + W_v H_{\text{image}}^{\text{attn}}), \quad (6)$$

$$H_{\text{fuse}} = (1-\lambda)\cdot H_{\text{news}} + \lambda \cdot H_{\text{image}}^{\text{attn}}, \quad (7)$$

Finally, the fused output H_{fuse} is inputted to the Transformer decoder to predict the target understanding Y.

B. Prompt tuning

In the task of prompt-based tuning, each input sample consisting of a pair $(\mathbf{x}_i, y_i)$ is transformed into a pattern-verbalizer pair (PVP) [26], denoted by $(p(\mathbf{x}_i), v(y_i))$. The pattern mapping function $p(\cdot)$ takes $\mathbf{x}_i$ as input and produces a cloze question with masked tokens. For instance, given a sentence represented as '$\mathbf{x}_i =$ [CLS] *News.* [SEP]*Correlation understanding.* [SEP]*Content understanding.* ', we can map it to a cloze question as follows: '$p\left(\mathbf{x}_i\right)$ = [CLS] *News.* The similarity score is *Correlation understanding.*The reasoning is *Content understanding.*The news is [MASK].[SEP]', where the tokens [CLS] and [SEP] serve as special start and end markers, respectively.

Within the context of prompt-based tuning, the function $v(\cdot)$, referred to as the verbalizer function, serves to map the label y_i to tokens that accurately represent its underlying semantic meaning. Given a PVP, the input representation $\mathbf{x}_i$ is derived by obtaining the token embedding $\mathbf{h}_i^{[\text{MASK}]}$ that corresponds to the [MASK] token. The class prediction $\hat{\mathbf{y}}_i$ is then generated as a probability distribution over all possible class labels, with the probability of the ground truth label y_i given $\mathbf{x}_i$ being estimated as:

$$q\left(y_i \mid \mathbf{x}_i\right) = \frac{\exp\left(\mathbf{w}_{v(y_i)}^\top \cdot \mathbf{h}_i^{[\text{MASK}]}\right)}{\sum_{y_j\in\mathcal{Y}} \exp\left(\mathbf{w}_{v(y_j)}^\top \cdot \mathbf{h}_i^{[\text{MASK}]}\right)} \quad (8)$$

Here, $\mathbf{w}_v$ denotes the logit vector of token v in the vocabulary, and $\mathcal{Y}$ refers to the set of all possible class labels. Let $\mathbf{y}_i$ be a one-hot vector with all elements being 0 except for the element corresponding to the ground truth class label $y_i \in 1, \ldots, C$. The model is trained using cross-entropy loss $\mathcal{L}_{\text{CE}}$, which is defined as follows:

$$\mathcal{L}_{\text{CE}} = \sum_{i=1}^{N} -\log\left(\hat{\mathbf{y}}_i\right)^\top \mathbf{y}_i, \quad (9)$$

where $(\cdot)^\top$ denotes the transpose operation, and N represents the total number of samples in the training set.

C. Contrastive Training

In order to enhance the discriminative power of news representation, we propose the implementation of a supervised contrastive learning objective, which serves to effectively cluster samples belonging to the same class while simultaneously separating those belonging to different classes:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_{y_i}-1}\sum_{j=1}^{N}\mathbb{1}_{[i\neq j]}\mathbb{1}_{[y_i=y_j]} log\frac{exp(\text{sim}(o_i,o_j)/\tau)}{\sum_{k=1}^{N}\mathbb{1}_{[i\neq k]}exp(\text{sim}(o_i,o_k)/\tau)} \quad (10)$$

where N_{y_i} is the number of news examples with the same label y_i, and $\mathbb{1}$ is the indicator. sim$(\cdot)$ denotes the cosine similarity function and τ controls the temperature.

Then we jointly train the model with the cross-entropy and supervised contrastive objectives:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{CL}}$$

where α is a hyperparameter to control the contribution of this $\mathcal{L}_{\text{Cl}}$. Algorithm 1 presents the training process of our approach.

IV. Experiments

In this section, we present the experiments designed to evaluate the effectiveness of MFDCoT. Our objective is to address the following evaluation questions: **EQ1**: To what extent can MFDCoT enhance the performance of few-shot MFD through the use of multimodal chain-of-thought reasoning? **EQ2**: How impactful are the benefits of multimodal chain-of-thought reasoning on the detection performance of MFDCoT?

Algorithm 1 Contrastive Prompt Tuning with Multimodal-CoT

Input: A small set of news C with news C_{news}, image C_{image}
Output: Assign news labels y to given unlabeled target data.

for each mini-batch $\mathcal{N}^i$ of the news C **do**:
  Generate news understanding $\mathcal{U}^i = F(\mathcal{N}^i_{\text{news}}, \mathcal{N}^i_{\text{image}})$ using the model $F(\cdot)$
  Generate correlation understanding $\mathcal{R}^i = P(\mathcal{N}^i_{\text{news}}, \mathcal{N}^i_{\text{image}})$ using the model $P(\cdot)$
  Pass $\mathcal{N}^i, \mathcal{U}^i, \mathcal{R}^i$ to the PVP and then PLM to obtain its [MASK] token representation $\mathbf{mask}^i$.
  Compute the classification loss $\mathcal{L}_{\text{CE}}$.
  Compute the supervised contrastive loss $\mathcal{L}_{\text{CL}}$.
  Jointly optimize all parameters of the model using the loss $\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{CL}}$.
procedure F(X)
  Encode the news and image inputs H_{news} and H_{image} via PLMs, respectively
  Build the interaction between news and news-related image features by attention $H^{\text{attn}}_{\text{image}}$
  Fuse H_{news} and $H^{\text{attn}}_{\text{image}}$ by a gated fusion mechanism to have H_{fuse}
  Feed H_{fuse} to the decoder to obtain the news understanding Y
  return Y
end procedure
procedure P(X)
  Encode the news and image inputs H_{news} and H_{image} via CLIP
  Compute to obtain the correlation understanding Y
  return Y
end procedure

EQ3: Can MFDCoT improve the accuracy of the zero-shot fake news detection task?

A. Datasets

To evaluate the efficacy of our approach, we performed experiments on two distinct datasets: PolitiFact and GossipCop. These datasets were sourced from The FakeNewsNet repository, with the PolitiFact dataset representing the political domain and the GossipCop dataset representing the entertainment domain [27]. Each news item in these datasets includes a full-length article as well as an associated image, and has been meticulously fact-checked by domain experts to ensure the accuracy of the labels assigned to them. In order to ensure a fair comparison with previous work, we applied the same dataset pre-processing methodology as utilized in Spotfake+ [5]. The key statistics pertaining to these datasets are presented in Table I.

Few-shot settings. In order to replicate low-resource conditions that are often encountered in real-world scenarios, we randomly sample $k \in (8, 64)$ instances as the training set. The development set and test set are kept at their original sizes. As the choice of training and development sets can have a significant impact on the test performance, we repeat this data sampling process using 10 different random seeds and report the average score obtained after removing the highest and lowest scores.

TABLE I
THE STATISTICS FOR THE TWO DATASETS.

Statistics	GossipCop	PolitiFact
# total news	12840	485
# fake news	2581	164
# real news	10259	321
# images per news	1	1

B. Baseline Model

We compare our proposed model against several state-of-the-art baseline methods, which are described as follows: 1)**TextCNN** [28]: This model is based on Convolutional Neural Networks (CNN) and is used for detecting misinformation by representing relevant news as a fixed-length sequence. 2)**LSTM-ATT** [29]: This is a model based on Long Short-Term Memory (LSTM) that uses an attention mechanism to consider the importance of words in the relevant news. 3)**FT+BERT** [30]: We use an existing fine-tuning technique based on the BERT pretrained language model. 4)**FakeBERT** [3]: This model combines different parallel blocks of single-layer deep CNNs with different kernel sizes and filters with the BERT model. 5)**SAFE** [6]: This model takes into account the relevance between news textual and visual information into a classifier to detect fake news. 6)**Spotfake+** [5]: This model uses VGG and BERT to respectively extract image and text features and concatenates them to detect fake news. 7)**CB-fake** [31]: This model uses Capsule Networks (CapsNet) and a pre-trained BERT model to capture more informative visual and textual features for fake news detection. 8)**CLIP-fake** [12]: This model utilizes the Contrastive Language-Image Pretraining (CLIP) model, which is pre-trained on large amounts of image and text data, for fake news detection. 9)**MFDCoT**: We improve an existing prompt-based tuning technique for fake news detection and extend it using multimodal chain-of-thought reasoning and contrastive learning. We evaluate the performance of these models in the most challenging setting of detecting fake news in few-shot scenarios.

C. Evaluation Metrics and Parameter Settings

Our evaluation metrics for this study, as in previous research [5], [12], comprised of accuracy and F1 score. To conduct our experiments, we utilized the pre-trained language model FlAN-T5 and improved upon the prompt-based tuning method for fake news detection. We employed the Adam optimizer, and selected the learning rate from the range of [1e-5 : 2e-5 : 5e-5], the hyperparameter α from [0 : 0.2 : 0.5 : 1], and the batch size from [2 : 4 : 8] for our MFDCoT model. To train our models, we conducted a maximum of 1000 steps and evaluated performance on the development set every 100 steps. All experiments were performed on a 32GB NVIDIA Tesla V100 GPU.

D. Results

To answer EQ1, we conducted a comparison of MFDCoT with the baselines described in Section IV-B for few-shot MFD. Table II presents the performance of our proposed method and

TABLE II
TEST PERFORMANCE (%) MEASURED ON FEW-SHOT FAKE NEWS DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	PolitiFact				GossipCop			
	8-shot		**64-shot**		**8-Shot**		**64-shot**	
	Acc.	Mac-F_1	Acc.	Mac-F_1	Acc.	Mac-F_1	Acc.	Mac-F_1
TextCNN	0.354	0.452	0.569	0.638	0.418	0.547	0.514	0.409
LSTM-ATT	0.377	0.481	0.584	0.657	0.462	0.521	0.520	0.512
FT+BERT	0.532	0.546	0.674	0.702	0.548	0.557	0.618	0.611
FakeBERT	0.542	0.585	0.695	0.773	0.588	0.578	0.638	0.680
SAFE	0.598	0.612	0.702	0.789	0.564	0.598	0.688	0.725
Spotfake+	0.615	0.638	0.725	0.811	0.612	0.665	0.676	0.743
CB-fake	0.621	0.643	0.721	0.813	0.628	0.686	0.683	0.761
CLIP-fake	0.673	0.683	0.737	0.828	0.653	0.711	0.692	0.788
MFDCoT	**0.762**	**0.817**	**0.815**	**0.891**	**0.725**	**0.798**	**0.767**	**0.836**

the compared methods on the GossipCop and PolitiFact test sets. From the table, several observations can be made:

(1) Methods that rely solely on unimodal features (TextCNN, LSTM-ATT, FT-BERT, and FakeBERT) exhibit poorer performance. This indicates that unimodal features are unable to effectively encode the semantic information of news content. Additionally, these methods are prone to overfitting in small sample scenarios, which makes it difficult for them to fully learn the differences between fake and true news.

(2) Many MFD methods, such as SAFE, CB-fake, and Spotfake+, rely solely on fused features obtained through concatenation or attention mechanisms. However, these fused features are not able to provide sufficient discrimination ability for classifying fake news. This is because the text and image features extracted separately are not in the same semantic space, and the correlation information between the text and image is not adequately captured during the fusion process.

(3) MFDCoT achieves a significant improvement in performance compared to CLIP-fake, mainly due to the following reasons. Firstly, the understanding news with images module in MFDCoT is able to generate semantically information-rich and news-related understanding in the same semantic space as multimodal pre-trained language models, reflecting the correlation between text and image and providing complementary information for fake news detection. The correlation understanding mechanism also determines the similarity score between news and news-related images, thereby avoiding the influence of invalid features on the representation ability of final features. Finally, the use of contrastive learning further improves classification accuracy.

E. Ablation Studies

To address EQ2, we conducted additional ablation studies on the various modules of MFDCoT. The compared variants of MFDCoT are implemented as follows:1)MFDCoT without CL: We remove the contrastive learning module. 2)MFDCoT without F: We remove the "understand news with images" module and use two unimodal features to classify news. 3)MFDCoT without C: We remove all CLIP-related modules.

TABLE III
ABLATION STUDY ON THE ARCHITECTURE DESIGN AND DIFFERENT FEATURES OF FNDCOT ON TWO DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	PolitiFact		GossipCop	
	64-shot		64-shot	
	Acc.	Mac-F_1	Acc.	Mac-F_1
MFDCoT image-only	0.652	0.689	0.588	0.625
MFDCoT text-only	0.782	0.811	0.746	0.793
MFDCoT w/o U	0.786	0.823	0.743	0.781
MFDCoT w/o C	0.807	0.838	0.752	0.785
MFDCoT w/o CL	0.802	0.861	0.748	0.782
MFDCoT	**0.815**	**0.891**	**0.767**	**0.836**

4)MFDCoT image-only: We remove all text-related features and only use the image feature extracted by DETR to classify. 5)MFDCoT text-only: We only use PLM-extracting feature to complete the detection task without any visual information.

Table III presents the experimental results obtained from two datasets. Our analysis initially focuses on assessing the impact of different modalities on the task of fake news detection. Notably, we observe a significant decline in performance across all datasets when news text is absent. This decline can be attributed to the increased difficulty of identifying fake news when relevant information is missing. By extracting and analyzing a broader range of modal information related to news, the model's detection performance can be significantly improved. Subsequently, we assess the impact of the "understand news with images" module. Our findings indicate a considerable decline in performance in the absence of the U and C components. The "understand news with images" module allows the exploitation of correlations between representations from different modalities through pre-trained language models (PLMs). This process facilitates fake news detection by providing a better understanding of the news content. Therefore, comprehending the news content is essential in effectively detecting fake news.

F. Zero-Shot Fake News Detection

Previous studies [2], [6], [9], [19] have shown that detecting fake news in minority domains is difficult due to the lack of

TABLE IV
ZERO-SHOT FAKE NEWS DETECTION RESULTS (%). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Target (Source)	PolitiFact (GossipCop)		GossipCop (PolitiFact)	
Model	Acc.	Mac-F_1	Acc.	Mac-F_1
FakeBERT	0.501	0.528	0.478	0.496
SAFE	0.639	0.694	0.576	0.673
CB-fake	0.643	0.682	0.581	0.677
CLIP-fake	0.661	0.718	0.594	0.693
MFDCoT	**0.782**	**0.818**	**0.693**	**0.754**

annotated resources. Additionally, unforeseen breaking events that were not covered in yesterday's news further exacerbate the scarcity of data resources. To address EQ3, we explored the possibility of utilizing multimodal chain-of-thought reasoning via domain transfer by employing the MFDCoT approach for zero-shot detection of fake news. Specifically, these models were trained on a source training set consisting of 64 examples and subsequently evaluated on the target test set.

Table IV displays the accuracy scores of several models, highlighting the superiority of MFDCoT over fine-tuning methods and state-of-the-art multimodal fusion techniques. The results demonstrate that MFDCoT attains an accuracy of approximately 78% on the PolitiFact dataset and 69% accuracy on the GossipCop dataset, significantly outperforming most of the baseline models. These experimental findings suggest that MFDCoT not only enhances few-shot detection performance, but also exhibits remarkable zero-shot fake news detection capabilities.

V. CONCLUSION

We formally study the problem of multimodal fake news detection. Our approach, MFDCoT, improves the detection of fake news by using a two-stage framework that combines multimodal chain-of-thought reasoning and contrastive prompting. In the first stage, we generate an understanding of news with images. Then in the second stage, we detect fake news based on this improved understanding. By incorporating a multimodal chain-of-thought reasoning module, our model achieves a thorough understanding of multimodal news and outperforms state-of-the-art models on both few-shot fake news classification and zero-shot detection tasks. We conducted extensive experiments on two real-world datasets, demonstrating the effectiveness of our proposed model.

REFERENCES

[1] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," *IJCAI*, 2016.

[2] Y. Long, Q. Lu, R. Xiang, M. Li, and C. Huang, "Fake news detection through multi-perspective speaker profiles," in *IJCNLP 2017, Volume 2: Short Papers*, pp. 252–256, 2017.

[3] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia tools and applications*, vol. 80, pp. 11765–11788, 2021.

[4] G. Jiang, S. Liu, Y. Zhao, Y. Sun, and M. Zhang, "Fake news detection via knowledgeable prompt learning," *Information Processing & Management*, vol. 59, no. 5, p. 103029, 2022.

[5] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *AAAI*, vol. 34, pp. 13915–13916, 2020.

[6] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," 2020.

[7] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," 2022.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[10] G. Bhatt, A. K. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "Combining neural, statistical and external features for fake news stance identification," *WCCP*, 2018.

[11] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," *Cornell University - arXiv*, 2020.

[12] Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang, "Multimodal fake news detection via clip-guided learning," *arXiv:2205.14304*, 2022.

[13] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," *The Web Conference*, 2019.

[14] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," 2022.

[15] H. Ye, N. Zhang, S. Deng, X. Chen, H. Chen, F. Xiong, X. Chen, and H. Chen, "Ontology-enhanced prompt-tuning for few-shot learning," 2023.

[16] K. Zhou, J. Yang, C. Loy, Z. Liu, and J. L. D. Z. M. S. D. Z. Z. L. M. D, "Conditional prompt learning for vision-language models," 2023.

[17] T. Shin, Y. Razeghi, R. L. Logan, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *Cornell University - arXiv*, 2020.

[18] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners.," *Cornell University - arXiv*, 2020.

[19] C. B. El Vaigh, T. Girault, C. Mallart, and D. H. Nguyen, "Detecting fake news conspiracies with multitask and prompt-based learning," in *MediaEval 2021-MediaEval Multimedia Evaluation benchmark. Workshop*, pp. 1–3, 2021.

[20] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," *Cornell University - arXiv*, 2020.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv: Learning*, 2020.

[22] M. Mohtarami, J. Glass, and P. Nakov, "Contrastive language adaptation for cross-lingual stance detection," *Empirical Methods in Natural Language Processing*, 2019.

[23] D. Zhang, F. Nan, X. Wei, S. Li, H. Zhu, K. McKeown, R. Nallapati, A. Arnold, and B. Xiang, "Supporting clustering with contrastive learning," *arXiv preprint arXiv:2103.12953*, 2021.

[24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv: Computer Vision and Pattern Recognition*, 2020.

[26] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," *ECACL*, 2021.

[27] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, 2020.

[28] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel textcnn model," *Neurocomputing*, vol. 363, pp. 366–374, 2019.

[29] N. Aloshban, "Act: Automatic fake news classification through self-attention," in *12th ACM Conference on Web Science*, pp. 115–124, 2020.

[30] A. Malakhov, A. Patruno, and S. Bocconi, "Fake news classification with bert.," in *MediaEval*, 2020.

[31] B. Palani, S. Elango, and V. Viswanathan K, "Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.