Team Triple-Check at Factify 2: Parameter-Efficient Large Foundation Models with Feature Representations for Multi-Modal Fact Verification

Wei-Wei Du[1], Hong-Wei Wu[1], Wei-Yao Wang[1] and Wen-Chih Peng[1]

[1]*Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan*

Abstract

Multi-modal fact verification has become an important but challenging issue on social media due to the mismatch between the text and images in the misinformation of news content, which has been addressed by considering cross-modalities to identify the veracity of the news in recent years. In this paper, we propose the **Pre-CoFactv2** framework with new parameter-efficient foundation models for modeling fine-grained text and input embeddings with lightening parameters, multi-modal multi-type fusion for not only capturing relations for the same and different modalities but also for different types (i.e., claim and document), and feature representations for explicitly providing metadata for each sample. In addition, we introduce a unified ensemble method to boost model performance by adjusting the importance of each trained model with not only the weights but also the powers. Extensive experiments show that Pre-CoFactv2 outperforms Pre-CoFact by a large margin and achieved new state-of-the-art results at the Factify challenge at AAAI 2023. We further illustrate model variations to verify the relative contributions of different components. Our team won the first prize (F1-score: 81.82%) and we made our code publicly available[1].

Keywords

Multi-modal fact verification, Parameter-efficient foundation models, Unified ensemble learning

1. Introduction

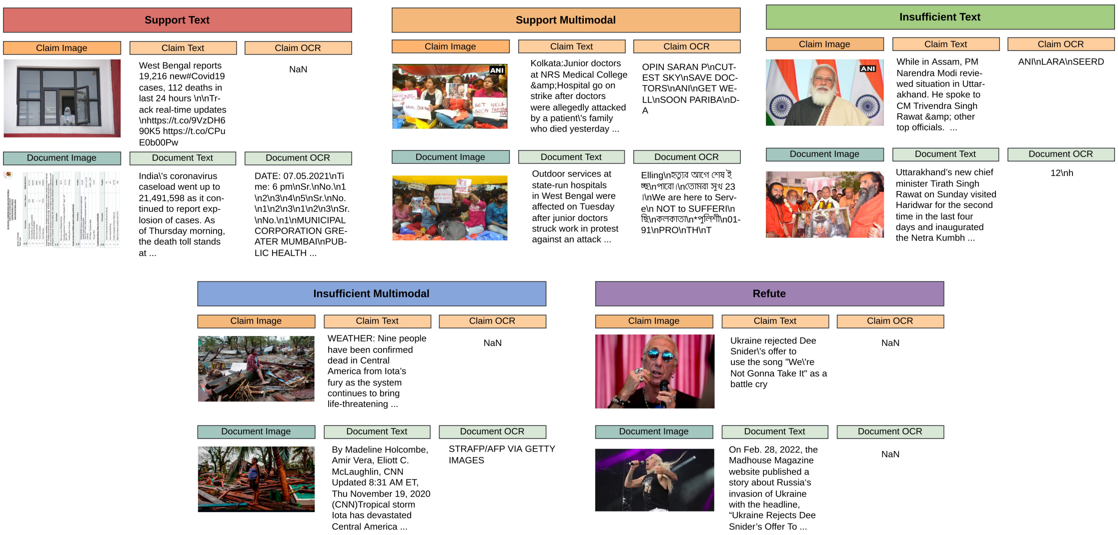
The rapid rise of social media technology allows people to send and receive information immediately, and also creates fertile soil for the fast spread of fake news. The proliferation of fake news not only triggers a storm of public opinion but also manipulates public events such as elections. For instance, there were approximately 30 million tweets from 2.2 million users on Twitter in the five months preceding the US 2016 presidential election, and either fake or extremely biased news was contained in 25% of these tweets [1], misleading people and seriously influencing the outcome of the election. Moreover, this issue became even worse during the COVID-19 pandemic period. Unconfirmed news with eye-catching images is becoming popular on social media since richer information easily attracts more viewers than news with only text. Therefore,

[1] https://github.com/wwweiwei/Pre-CoFactv2-AAAI-2023.

De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023. 2023 Washington, DC, USA

wwdu.cs10@nycu.edu.tw (W. Du); johnnyhwu.cs11@nycu.edu.tw (H. Wu); sf1638.cs05@nctu.edu.tw (W. Wang); wcpeng@cs.nycu.edu.tw (W. Peng)

CEUR Workshop Proceedings (CEUR-WS.org)



West Bengal reports 19,216 new#Covid19 cases, 112 deaths in last 24 hours \n\nTrack real-time updates \nhttps://t.co/9VzDH690K5 https://t.co/CPuE0b00Pw

NaN

India\'s coronavirus caseload went up to 21,491,598 as it continued to report explosion of cases. As of Thursday morning, the death toll stands at ...

DATE: 07.05.2021\nTime: 6 pm\nSr.\nNo.\n1\n2\n3\n4\n5\nSr.\nNo.\n1\n2\n3\n1\n2\n3\nSr.\nNo.\n1\nMUNICIPAL CORPORATION GREATER MUMBAI\nPUBLIC HEALTH ...

Kolkata:Junior doctors at NRS Medical College &amp;Hospital go on strike after doctors were allegedly attacked by a patient\'s family who died yesterday ...

OPIN SARAN P\nCUTEST SKY\nSAVE DOCTORS\nANI\nGET WELL\nSOON PARIBA\nDA

Outdoor services at state-run hospitals in West Bengal were affected on Tuesday after junior doctors struck work in protest against an attack ...

Elling\nহত্যার আগে শেষ ই চ্ছ\nপারো।\nতোমরা সুখ 23 ।\nWe are here to Serve\n NOT to SUFFERI\n ছি\nকলকাতা\n*পুলিশী\n01-91\nPRO\nTH\nT

While in Assam, PM Narendra Modi reviewed situation in Uttarakhand. He spoke to CM Trivendra Singh Rawat &amp; other top officials. ...

ANI\nLARA\nSEERD

Uttarakhand's new chief minister Tirath Singh Rawat on Sunday visited Haridwar for the second time in the last four days and inaugurated the Netra Kumbh ...

12\nh

WEATHER: Nine people have been confirmed dead in Central America from Iota's fury as the system continues to bring life-threatening ...

NaN

By Madeline Holcombe, Amir Vera, Eliott C. McLaughlin, CNN Updated 8:31 AM ET, Thu November 19, 2020 (CNN)Tropical storm Iota has devastated Central America ...

STRAFP/AFP VIA GETTY IMAGES

Ukraine rejected Dee Snider\'s offer to use the song "We\'re Not Gonna Take It" as a battle cry

NaN

On Feb. 28, 2022, the Madhouse Magazine website published a story about Russia's invasion of Ukraine with the headline, "Ukraine Rejects Dee Snider's Offer To ...

NaN

Figure 1: An example of each category in the Factify 2 dataset [6]. Each sample contains a claim and document, both of which are composed of an image, text, and OCR of the image.

in order to mitigate the negative impact caused by fake news, there is an urgent need to develop a multi-modal fake checker that can automatically assess the validity of news.

Given a claim and the support information consisting of not only text but also images, the fact checker aims to discriminate whether the claim entails the support. Recent approaches have demonstrated that utilizing multi-modal contexts to detect fake news achieves better performance than only using one modality. For instance, Wang et al. [2] learned event-invariant features using an adversarial network along with a multi-modal feature extractor to enhance the performance of the fake news detector. More recently, Wang and Peng [3] introduced Pre-CoFact with DeBERTa [4] and DeiT [5], to extract features from both claims and documents' text and images, respectively and then fuses multi-modal contexts with the co-attention modules, illustrating the competitive performance without auxiliary information.

However, Pre-CoFact fails to finetune the pre-trained model due to the large number of parameters, which heavily depend on the post layers to learn the contextual information. Besides, we argue that additional textual features such as stopword and URL counts provide definitive descriptions and relations between inputs. In addition, the co-attention modules only capture the dependencies between modalities while ignoring the relations between different types of samples (i.e., claim and document). Therefore, we believe that multi-modal fact verification remains an unexplored but essential problem.

To tackle this task and the aforementioned limitations, we examine our proposed method on the real-world Factify 2 challenge [6], which is the largest multi-modal fact verification dataset consisting of 50K news items from India and the US. Specifically, each sample contains one *document* with a related image representing the reliable source of information and one *claim* which also includes an associated image representing another source of information whose validity needs to be assessed. The training set, validation set, and testing set are composed of 35000 samples, 7500 samples, and 7500 samples, respectively. As shown in Figure 1, each sample contains a textual claim, claim image, optical character recognition (OCR) of the claim image, document, document image, and OCR of a document image, and is classified into support, insufficient evidence, and refute between given claims and documents (detailed labels will be introduced in §3.1)

In this paper, we propose **Pre-CoFactv2** with parameter-efficient large foundation models with feature representations to address the challenge. We utilize large-scale pre-trained foundation models, DeBERTa [4] and Swin Transformer v2 (Swinv2) [7], to extract contextualized embeddings from both textual content and visual images, respectively, and an adapter module is used to improve model performance by finetuning the backbone with only a few parameters. Furthermore, additional feature metadata such as the number of stopwords and URLs in the claim and document is generated from the input to enrich the information; this enables the model to learn explicit information from different perspectives. Afterwards, we integrate the information among several modalities with multi-modal multi-type fusion modules into corresponding embeddings to classify the category of the news.

To summarize, the contributions of this paper are three-fold:

- We propose parameter-efficient large foundation models with feature representations (Pre-CoFactv2) for multi-modal fact verification by integrating adapters in the large-scale foundation models to achieve competitive performance by training only lightening parameters, and converting additional features to learn the explicit correlations between claim and document.
- In addition to capturing information between modalities, we design multi-modal multi-type fusions for different types of modalities (i.e., claim text and document images, document text and claim images), which enforces the model to distinguish between the given claim and document.
- To boost the detection quality, we introduce a unified ensemble method to integrate various considerations from diverse models. This approach won first place, surpassing the second place by 1.3% and the official baseline by 25.9% in terms of testing score. Moreover, extensive experiments were conducted to examine and analyze the contribution and effectiveness of each module.

2. Related Works

2.1. Multi-Modal Fact Verification

In recent years, multiple modalities (e.g., text and images) have been incorporated to demonstrate the great potential for fact verification. MAVE [8] was proposed to learn better multi-modal shared representations with a variational autoencoder by jointly training with a fact verification classifier to verify the posts. In addition to learning shared features among several modalities, Qian et al. [9] introduced HMCAN with a multi-modal contextual attention module to model the features from multiple modalities in each news post, and a hierarchical encoding module to capture the rich hierarchical semantics of text. Besides, Wu et al. [10] proposed MCAN to take inter-modality relations into consideration by using multiple co-attention layers to fuse visual and textual features extracted from Transformer-based models. Recently, Chen et al. [11]

proposed CAFE, which consists of an alignment module to transform the features from each modality into a shared semantic space, an ambiguity module to estimate the ambiguity between different modalities and a fusion module to capture the multi-modal correlations.

Existing work focuses on the importance of learning shared representations among multi-modal information, which motivates us to use attention mechanisms to achieve the purpose. However, pre-trained models with a base size are often used in the previous approaches, which cannot utilize more fine-grained features compared with the large-size model. To that end, we adopt large-size pre-trained models to effectively produce embeddings of text and images from the complex multi-modal inputs.

2.2. Pre-Trained Model for Different Modalities

Transformer [12] has become the widely-used neural network architecture in various NLP and CV tasks due to its parallel computation and long-term considerations. Since the advent of BERT [13], a line of large-scale Transformer-based pre-trained language models (PLMs) such as GPT-3 [14], DeBERTa [4], PaLM [15], and BLOOM [16] have been introduced to demonstrate the generalizability of large-scale PLMs. With the drastically increased capacity of the model, the accuracy of various language benchmarks has been significantly improved; it was therefore been adopted to finetune downstream tasks and has achieved better performance.

In addition to the success of PLMs in NLP tasks, Transformer has also started taking over visual benchmarks recently. Dosovitskiy et al. [17] proposed Vision Transformer (ViT) for pre-training with image patches, which has achieved competitive results compared to state-of-the-art convolution neural networks on ImageNet-1K image-level classification benchmarks. Swin Transformer (Swinv1) [18] constructs hierarchical feature maps and uses the shifted window approach for computing self-attention, making it suitable as a general-purpose backbone for various vision tasks such as object detection and semantic segmentation. Afterwards, Swin Transformer v2 (Swinv2) [7] was proposed with several adaptations in order to better scale up model capacity and window resolution to mitigate the unstable training and size discrepancy between the pre-training and training images of Swinv1.

To use the generic knowledge of textual and visual information for fact verification, we employ SOTA large-scale PLMs as pre-trained models instead of learning from scratch. Furthermore, we utilize adapters in large-scale PLMs to finetune PLMs with lightening parameters while improving model performance.

3. Method

3.1. Problem Formulation

Given a multi-modal claim denoted by $C = \{C_i^T, C_i^I\}_{i=1}^{|C|}$ and a fact-checking document denoted by $D = \{D_i^T, D_i^I\}_{i=1}^{|D|}$, the goal is to classify one of the five categories:

- Support_Text: the claim text is similar but images of the document and claim are not similar.
- Support_Multimodal: both the claim text and image are similar to that of the document.

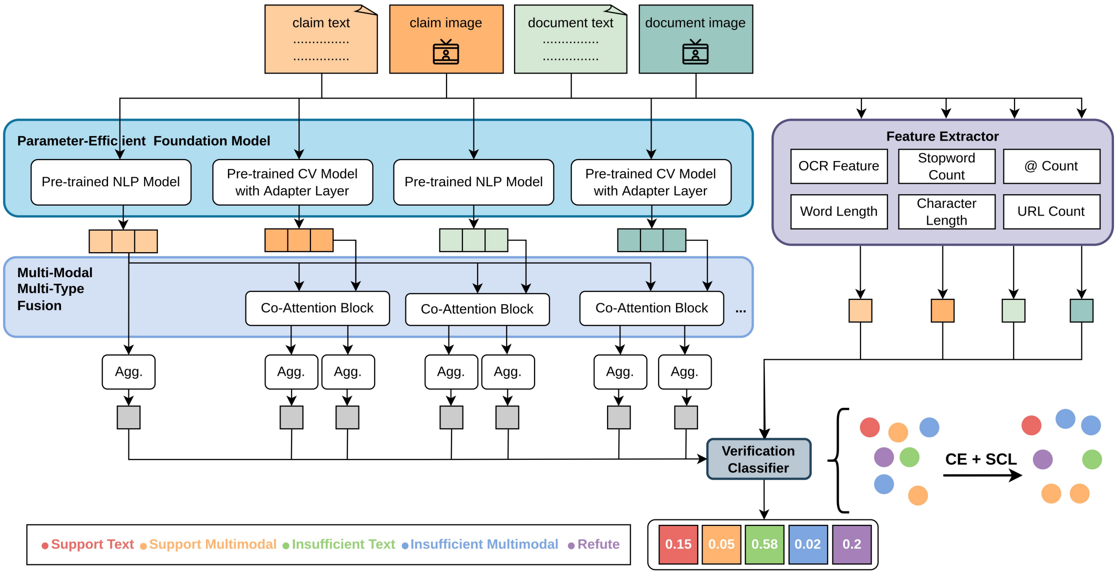


Figure 2: Illustration of the Pre-CoFactv2 framework. The parameter-efficient foundation model aims to transform the input text and images into embedding by the pre-trained language model. Then, the multi-modal multi-type fusion fuses the information from the same modality (images/text from the claim and document), different modalities (images and text from the claim/document), and different types (image from the claim and text from the document, and text from the claim and image from the document) to obtain contexts. Besides, the feature extractor is designed to convert input text and images into several features. In the end, the verification classifier contains cross-entropy loss and contrastive loss to predict the possible class based on the embeddings from previous outputs.

- Insufficient_Text: both text and images of the claim are neither supported nor refuted by the document.
- Insufficient_Multimodal: the claim text is neither supported nor refuted by the document but images are similar to the document.
- Refute: The images or text from the claim and document are completely contradictory.

Each sample contains a claim and a document, each of which includes the OCR feature, some special token count (e.g., stopword, @ URL), and word and character length.

3.2. Pre-CoFactv2 Overview

Figure 2 illustrates an overview of the proposed Pre-CoFactv2 framework. The additional features are generated by the feature extractor from the given claim text, claim image, document text, and document image. Then, we adopt two parameter-efficient foundation models for learning in-domain knowledge from pre-trained embeddings with adapters and a multi-modal multi-type fusion module for modeling not only cross-modality (i.e., text and image) relations but also cross-type (i.e., claim and document) relations. Outputs of these embeddings are fused by the verification classifier with cross-entropy loss as well as supervised contrastive loss [19] to separate embeddings and find clearer boundaries.

3.3. Feature Extractor

Inspired by previous works for textual information in fake checking [20, 21], textual features are extracted from both claim and document text to enable the model to learn explicit information from different perspectives. Specifically, we extract statistical features from the text, namely sentence word length, and character length. Then, we calculate the stopword, @name, and URL counts to indicate the style of text content. Besides, OCR text extracted from images is adopted to get the semantics information of the image instead of pixel values only.

3.4. Parameter-Efficient Foundation Model

3.4.1. Foundation Model

Benefiting from the advancement of utilizing basic knowledge as pre-trained models, we follow [3] to incorporate NLP and CV foundation models as the initialized embeddings of text and images. However, previous work failed to fully exploit the knowledge of generic datasets; we, therefore, adopt state-of-the-art foundation models with larger sizes as the pre-trained models. Specifically, we first use DeBERTa large [4] as our pre-trained NLP model and Swinv2 base [7] as our pre-trained CV model, and then the embedding layer is used for transforming pre-trained embeddings to embeddings in our task. Formally, the i-th output of the embedding layer is calculated as follows:

$$E_{C_i^I} = \sigma(W_{C^I}X_{C_i^I} + b_{C^I}); X_{C_i^I} = Swinv2(C_i^I), \tag{1}$$

$$E_{D_i^I} = \sigma(W_{D^I}X_{D_i^I} + b_{D^I}); X_{D_i^I} = Swinv2(D_i^I), \tag{2}$$

$$E_{C_i^T} = \sigma(W_{C^T}X_{C_i^T} + b_{C^T}); X_{C_i^T} = DeBERTa(C_i^T), \tag{3}$$

$$E_{D_i^T} = \sigma(W_{D^T}X_{D_i^T} + b_{D^T}); X_{D_i^T} = DeBERTa(D_i^T), \tag{4}$$

where the dimensions of $E_{C_i^I}, E_{D_i^I}, E_{C_i^T}, E_{D_i^T}$ are d, the activation function σ uses ReLU [22].

3.4.2. Parameter-Efficient Adapter

Finetuning existing foundation models requires a large number of computation resources due to the number of parameters, but is able to learn better representations for downstream tasks which are not trained in Pre-CoFact. To that end, we design an adapter module in the foundation models, which enables us to finetune the backbone with only a few parameters but still achieves better performance than the freeze parameters.

We follow [23], which demonstrates the competitive performance of fine-tuning between lightening parameters and all backbone parameters, by adding an additional adapter layer in the output layer of Swinv2 to empower the model to capture the information of images in the downstream tasks (fake news detection in this paper). The weights of the adapter layer are computed as follows:

$$Swinv2(X) = FFN(\tilde{X}) + Adapter(\tilde{X}); Adapter(\tilde{X}) = (W\tilde{X} + b) + v, \tag{5}$$

where $\tilde{X}$ is the output embedding before the feed-forward layer of Swin-Transformerv2, $FFN(X)$ is the original feed-forward layer, and $Adapter(X)$ produces the adapted representations with the same dimension of $FFN(X)$ from the input-related weights.

We note that we cannot add adapter layers to our NLP foundation models since our GPU memory will OOM, but we believe that this concept can improve the performance as well.

3.5. Multi-Modal Multi-Type Fusion

Previous work only considered 1) images of claims and documents, 2) text of claims and documents, 3) images of claims and text of documents, and 4) images of claims and text of claims to produce context embeddings. However, the relations between different types are critical to judge the text and images across claims and documents. Therefore, we add two additional co-attention blocks to model correlations between 5) images of documents and text of claims, and 6) images of documents and text of documents. The co-attention block is a variation of the multi-head self-attention block, which takes two modalities as inputs to learn interactions and relations. Specifically, we compute the dot products of the query and the key, divide each by $\sqrt{d}$, and apply a softmax function to the attention scores to obtain the weights on the values, which indicates the relative importance of each value for a given query.

We illustrate the computation of images of claims (E_{C^I}) and documents (E_{D^I}) as examples, and the others follow a similar process:

$$Q_C^I = E_{C^I} W^{Q_C^I}, K_C^I = E_{C^I} W^{K_C^I}, V_C^I = E_{C^I} W^{V_C^I}, \tag{6}$$

$$Q_D^I = E_{D^I} W^{Q_D^I}, K_D^I = E_{D^I} W^{K_D^I}, V_D^I = E_{D^I} W^{V_D^I} \tag{7}$$

$$Att(E_{C^I}, E_{D^I}) = softmax(\frac{Q_C^I (K_D^I)^T}{\sqrt{d}}) V_D^I) \tag{8}$$

$$Z^I = Norm(E_{C^I} + Att(E_{C^I}, E_{D^I})) \tag{9}$$

$$O_{C,D}^I = Norm(FFN(Z^I) + Z^I) \tag{10}$$

where $W^{Q_C}, W^{K_C}, W^{V_C}, W^{Q_D}, W^{K_D}, W^{V_D} \in \mathbb{R}^{d \times d}$, and $Norm$ and FFN is the same normalization method and feed-forward network as in [12].

To use fewer parameters, we share weights in a co-attention block for improving performance because of allowing the model to learn common representations of the inputs. Finally, we apply the mean aggregation to fuse all the results into one embedding to represent the corresponding sentence or image embeddings.

3.6. Category Classifier

For predicting the label of the given claims and documents, all of the 12 aggregated outputs from the multi-modal multi-type fusion, the 4 aggregated embeddings of $E_{C_i^I}, E_{D_i^I}, E_{C_i^T}, E_{D_i^T}$, and the 32 dimensions output from the feature extractor are concatenated as the input O of the classifier:

$$\hat{y}_i = softmax((\sigma(OW^{Z1}))W^{Z2}), \tag{11}$$

where $W^{Z1} \in \mathbb{R}^{48d \times d_m}$ and $W^{Z2} \in \mathbb{R}^{d_m \times 5}$. Note that σ uses ReLU which is the same as E.

To enhance the generalization of our method, we jointly train supervised contrastive learning [19] and cross-entropy. Thus, embeddings with the same label would become closer, while embeddings with different labels would increase the distance. The loss function is as follows:

$$\mathbb{L} = \alpha \times - \sum_{i=1}^{|C|} y_i log(\hat{y_i}) + (1 - \alpha) \times SupConLoss, \tag{12}$$

where *SupConLoss* indicates a supervised contrastive loss. We set 0.7 for cross-entropy loss and 0.3 for supervised contrastive loss respectively[1].

3.7. Unified Ensemble Techniques

To eliminate the effect of noisy data and to integrate various advantages, ensemble learning with power weighted sum has been used to integrate the informative knowledge from different models to achieve a better predictive performance of the overall model via the voting technique [3, 24]. However, we argue that using different powers for each model improves the generalizability since each model does not require to use of the same projection space. Therefore, we propose a unified ensemble method that the final predicted probabilities P are computed with the independent power weighted sum as:

$$P = P_1^{N_1} \times w_1 + \cdots + P_M^{N_M} \times w_M, \tag{13}$$

where M is set to 3 in this work, $w1, \cdots, w_M$ are weights of the corresponding model, and $N_1, \cdots, N_M$ are weights of power. We tune these hyper-parameters based on the validation set and use ensemble weights as 0.2, 0.7, and 0.6 and powers as 0.125, 0.125, and 0.25, respectively.

4. Results and Analysis

4.1. Implementation Details

The dimension of additional features from text and image was set to 32, the embedding dimension d of text and image were both set to 256, the inner dimension of the feed-forward layer was 512, and the number of heads was set to 12. The hidden dimension of the classifier d_m was 128. The dropout rate was 0.1, and the max sequence length was 512. The batch size was 24, the learning rates were set to 5e-5 and 1e-5, the pre-training epochs were set to 10 and the training epochs were set to 15, and the seeds were tested with 42. The pre-trained DeBERTa was deberta-large[2], and the Swinv2 was swinv2-base-patch4-window8-256[3]. All of the parameters in the two pre-trained models are first finetuned by another dataset [25] to enhance the model capability, and then by the Factify 2 dataset. In the feature extractor module, all images were transformed by resizing to 256, center cropping to 256, and normalizing. On the other hand, all text was normalized by replacing emojis in text strings using demoji [4], all the abbreviations are

[1] We empirically found that supervised contrastive loss is not beneficial in this task (§4.2.1), thus the final model sets α as 1.

[2] https://huggingface.co/microsoft/deberta-large

[3] https://huggingface.co/microsoft/swinv2-base-patch4-window8-256

[4] https://pypi.org/project/demoji/

Module Used	Pre-CoFact	(1)	(2)	(3)	(4)	(5)
Feature Extractor	x	v	v	x	v	x
PLM_{CV}	DT-b	CN	SW-b	SW-b	SW-b	SW-b
PLM_{NLP}	DE-b	BB	DE-l	DE-l	DE-l	DE-l
Adapters	x	x	x	x	v	v
SupConLoss	x	x	x	x	v	x
Weighted F1 (%)	74.22	74.67	76.60	75.89	75.56	**78.80**

Table 1
Variations of our model by validation score. The pre-trained models include: DeiT-base (DT-b), CoatNet (CN), Swinv2-base (SW-b), Deberta-base (DE-b), Deberta-large (DE-l), BigBird (BB).

expanded, and we delete the @name and URL to shorten the length of the text with meaningless words. All the experiments were conducted on a machine with AMD Ryzen Threadripper 3960X 24-Core Processor, Nvidia GeForce RTX 3090, and 252GB RAM. To evaluate the performance of the task, the weighted average F1 score was used across the 5 categories. The source code is available at https://github.com/wwweiwei/Pre-CoFactv2-AAAI-2023.

4.2. Ablation Study

4.2.1. Variations of Our Proposed Model

To examine the relative contribution of our proposed module, we first conduct the ablative experiments by removing each module as the variants of Pre-CoFactv2. Table 1 summarizes the results of the variations on the validation set. We can observe that adopting different pre-trained models (i.e., CoatNet [26] and BigBird [27]) with feature extractor (1) slightly outperforms Pre-CoFact. Moreover, using a pre-trained CV model with Swinv2 with an adapter and a pre-trained NLP model with Deberta-large (3) significantly improves the performance compared with Pre-CoFact, which illustrates the importance of not only larger PLMs and finetuning them with lightening parameters but also the multi-modal multi-type fusion. Comparing the variations (2) with (3), additional information from the feature extractor further boosts the performance by capturing explicit information instead of only semantic information from the embeddings. It is worth noting that adding contrastive loss to the embeddings after fusion fails to benefit overall performance, which suggests that calculating the contrastive loss of images and text respectively may be more effective. Nonetheless, our proposed framework signifies the ability to incorporate multi-modal claims and documents with explicit features and lightening parameters to effectively classify the veracity of the news.

4.2.2. Variations of Our Ensemble Technique

To ensure the effectiveness of our unified ensemble technique, we conducted a comprehensive ablation study of variants of ensemble techniques by integrating all the predictions in different weights and powers. (1) is the subset of (2) when all the weight is the same, (2) is the subset of

Model	(1) Average	(2) Weighted sum	(3) Power weighted sum	(4) Unified power weighted sum (Our)
Support Text F1 (%)	71.64	72.65	**73.50**	73.46
Support Multimodal F1 (%)	80.00	81.39	81.66	**81.88**
Insufficient Text F1 (%)	75.21	75.78	**76.39**	76.30
Insufficient Multimodal F1 (%)	74.62	76.22	76.11	**76.41**
Refute F1 (%)	99.73	99.70	**99.80**	**99.80**
Weighted F1 (%)	80.24	81.15	81.49	**81.57**

Table 2
Ablation study of ensemble techniques in terms of validation score. (1) Average: (prob1+prob2+prob3)/3, (2) Weighted sum: $w1 \times prob1 + w2 \times prob2 + w3 \times prob3$, (3) Power weighted sum: $w1 \times prob1^{p} + w2 \times prob2^{p} + w3 \times prob3^{p}$, (4) Unified power weighted sum (Ours): $w1 \times prob1^{p1} + w2 \times prob2^{p2} + w3 \times prob3^{p3}$.

(3) when all the power equals 1, and (3) is the subset of (4) when all the power is equal. Table 2 proves that the unified power weighted sum achieves the best result, which demonstrates that our proposed unified ensemble method is superior to other techniques by up to 1.7%.

We can observe that averaging all predictions improves the performance only slightly, while setting different importance and powers boosts the performance more. Our unified ensemble technique, in contrast, achieves the best quality compared to different ensemble methods. We note however that these parameters are manually tuned, which requires more costs to achieve better performance.

4.3. Testing Performance

The results for the testing set are shown in Table 3 in terms of the weighted F1-score. Our approach achieved the state-of-the-art performance of 81.82% of the weighted F1-score, winning first place in the multi-modal fact-checking challenge, and outperformed the second place and the official baseline by 1.3% and 25.9%, respectively. This again indicates the reasonable and effective design of Pre-CoFactv2.

To further analyze the classification details, we illustrate the confusion matrices of the validation set as well as the testing set of Pre-CoFactv2. As shown in Figure 3, we can observe that the class of refute is the most distinguishable category, while some of the classes between insufficient_text and support_text misclassify each other.

5. Conclusion

In this work, we introduce parameter-efficient large foundation models with feature representation (Pre-CoFactv2) to mitigate the issue of disseminating multi-modal fake news. With the integration of adapters with foundation models, we are able to finetune the backbones with only a few parameters while achieving better performance. In addition, feature representations provide explicit information about both text and images of claims as well as documents to clarify

	Triple-Check	Baseline
Support_Text (%)	**82.77**	50.00
Support_Multimodal (%)	**91.38**	82.72
Insufficient_Text (%)	**85.19**	80.24
Insufficient_Multimodal (%)	**89.22**	75.93
Refute (%)	**100**	98.82
Final (%)	**81.82**	64.99

Table 3
Performance of Pre-CoFactv2 in terms of testing score, which achieved the first place and outperformed the official baseline [28] by 25.9%.

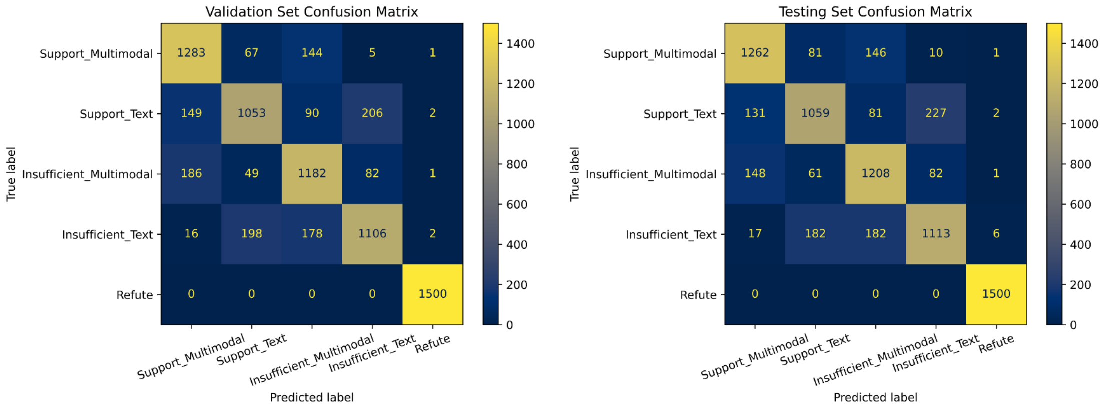


Figure 3: The confusion matrices of the validation set and testing set of our proposed model Pre-CoFactv2.

the relations between them with multi-modal multi-type fusion. With the help of a unified ensemble technique, Pre-CoFactv2 is ranked first on the official leaderboard with an F1 score of 0.81, which outperforms the baseline by 25.9% and greatly benefits the research of multi-modal fact verification. Furthermore, extensive ablative experiments demonstrate the effectiveness of each module in our proposed framework.

References

[1] A. Bovet, H. A. Makse, Influence of fake news in twitter during the 2016 US presidential election, CoRR abs/1803.08491 (2018).

[2] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, EANN: event adversarial neural networks for multi-modal fake news detection, in: KDD, ACM, 2018, pp. 849–857.

[3] W. Wang, W. Peng, Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification (short paper), in: DE-FACTIFY@AAAI, volume 3199 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[4] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: ICLR, OpenReview.net, 2021.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: ICML, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10347–10357.

[6] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Factify 2: A multimodal fake news and satire news dataset, in: proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.

[7] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer V2: scaling up capacity and resolution, in: CVPR, IEEE, 2022, pp. 11999–12009.

[8] D. Khattar, J. S. Goud, M. Gupta, V. Varma, MVAE: multimodal variational autoencoder for fake news detection, in: WWW, ACM, 2019, pp. 2915–2921.

[9] S. Qian, J. Wang, J. Hu, Q. Fang, C. Xu, Hierarchical multi-modal contextual attention network for fake news detection, in: SIGIR, ACM, 2021, pp. 153–162.

[10] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: ACL/IJCNLP (Findings), volume ACL/IJCNLP 2021 of *Findings of ACL*, Association for Computational Linguistics, 2021, pp. 2560–2569.

[11] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: WWW, ACM, 2022, pp. 2897–2905.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: NeurIPS, 2020.

[15] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, CoRR abs/2204.02311 (2022).

[16] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-

Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, et al., BLOOM: A 176b-parameter open-access multilingual language model, CoRR abs/2211.05100 (2022).

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, OpenReview.net, 2021.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: ICCV, IEEE, 2021, pp. 9992–10002.

[19] B. Gunel, J. Du, A. Conneau, V. Stoyanov, Supervised contrastive learning for pre-trained language model fine-tuning, in: ICLR, OpenReview.net, 2021.

[20] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: WWW, ACM, 2011, pp. 675–684.

[21] J. Gao, H. Hoffmann, S. Oikonomou, D. Kiskovski, A. Bandhakavi, Logically at the factify 2022: Multimodal fact verification, CoRR abs/2112.09253 (2021).

[22] A. F. Agarap, Deep learning using rectified linear units (relu), CoRR abs/1803.08375 (2018).

[23] C. Fu, Z. Chen, Y. Lee, H. Lee, Adapterbias: Parameter-efficient token-dependent representation shift for adapters in NLP tasks, in: NAACL-HLT (Findings), Association for Computational Linguistics, 2022, pp. 2608–2621.

[24] W. Wang, Y. Tang, W. Du, W. Peng, Nycu_twd@lt-edi-acl2022: Ensemble models with VADER and contrastive learning for detecting signs of depression from social media, in: LT-EDI, Association for Computational Linguistics, 2022, pp. 136–139.

[25] S. Mishra, S. S, A. Bhaskar, P. Chopra, A. N. Reganti, P. Patwa, A. Das, T. Chakraborty, A. P. Sheth, A. Ekbal, FACTIFY: A multi-modal fact verification dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, volume 3199 of *CEUR Workshop Proceedings*, ceur, 2022.

[26] Z. Dai, H. Liu, Q. V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, in: NeurIPS, 2021, pp. 3965–3977.

[27] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, in: NeurIPS, 2020.

[28] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Findings of factify 2: multimodal fake news detection, in: proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.