Team gzw at Factify 2: Multimodal Attention and Fusion Networks for Multi-Modal Fact Verification

Zhenwei Gao[1], Tong Chen[1] and Zheng Wang[1,2,*]

[1]Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, China

[2]Institute of Electronic and Information Engineering of UESTC in Guangdong, 523808

Abstract

Nowadays, detecting fake news on social media platforms has become a top priority since the widespread dissemination of fake news may mislead readers and have negative effects. To address the problem, we propose a Multimodal Attention and Fusion Network (MAFN) for multi-modal fact verification. Specifically, we employ DeiT and DeBERTa to obtain better representations for text and images, respectively. Then, we feed the obtained representations of images and text into a multi-modal attention network to fuse both inter-modality and intra-modality relationships. Besides, we adopt an ensemble strategy by using different pre-trained models in MAFN to achieve better performance. We conduct a series of ablation studies to verify the impact of each designed module on performance. Our method (team gzw) ranked fifth in the leaderboard of the Factify Challenge hosted by De-Factify@AAAI 2023, achieving an F1 score of 76.051%, which shows that our model achieves a competitive performance.

Keywords

Multi-modal Attention, Pre-trained Model, Self-Attention, De-Factify

1. Introduction

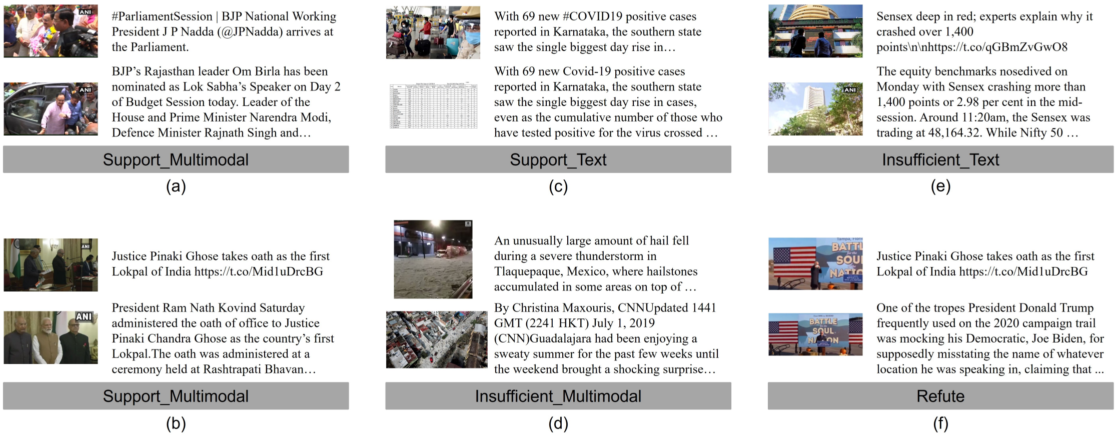
Social media has become a mainstream platform for people to communicate their ideas, due to the increasing convenience and intelligence. However, every coin has two sides. That is to say, it also gradually becomes an ideal place for the widespread of fake news. Since fake news distorts and fabricates facts maliciously, its extensive dissemination has extremely negative impacts on individuals and society. In addition, multimedia intelligence [1, 2, 3] can help the people better understand the world. Therefore, it is urgently important to detect fake news with multimedia in social platforms.

In order to facilitate the detection of fake news, many approaches have been proposed. The early attempts (e.g., snopes.com) mainly verified the fake news by experts or institutions in related fields, which is obviously time-consuming and labor-intensive. Therefore, automatically detecting fake news has been a key research direction and drawn much attention in recent years. Basically, existing studies on automatic fake news detection can be summarized into two

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

[*]Corresponding author. This work was supported by the Sichuan Science and Technology Program (23ZDYF0618) and Guangdong Basic and Applied Basic Research Foundation (2022A1515110576).

gaozhenwei69@gmail.com (Z. Gao); chentonglucky6@163.com (T. Chen); zh_wang@hotmail.com (Z. Wang)



#ParliamentSession | BJP National Working President J P Nadda (@JPNadda) arrives at the Parliament.

BJP’s Rajasthan leader Om Birla has been nominated as Lok Sabha’s Speaker on Day 2 of Budget Session today. Leader of the House and Prime Minister Narendra Modi, Defence Minister Rajnath Singh and…

Justice Pinaki Ghose takes oath as the first Lokpal of India https://t.co/Mid1uDrcBG

President Ram Nath Kovind Saturday administered the oath of office to Justice Pinaki Chandra Ghose as the country’s first Lokpal.The oath was administered at a ceremony held at Rashtrapati Bhavan…

With 69 new #COVID19 positive cases reported in Karnataka, the southern state saw the single biggest day rise in…

With 69 new Covid-19 positive cases reported in Karnataka, the southern state saw the single biggest day rise in cases, even as the cumulative number of those who have tested positive for the virus crossed …

An unusually large amount of hail fell during a severe thunderstorm in Tlaquepaque, Mexico, where hailstones accumulated in some areas on top of …

By Christina Maxouris, CNNUpdated 1441 GMT (2241 HKT) July 1, 2019 (CNN)Guadalajara had been enjoying a sweaty summer for the past few weeks until the weekend brought a shocking surprise…

Sensex deep in red; experts explain why it crashed over 1,400 points\n\nhttps://t.co/qGBmZvGwO8

The equity benchmarks nosedived on Monday with Sensex crashing more than 1,400 points or 2.98 per cent in the mid-session. Around 11:20am, the Sensex was trading at 48,164.32. While Nifty 50 …

Justice Pinaki Ghose takes oath as the first Lokpal of India https://t.co/Mid1uDrcBG

One of the tropes President Donald Trump frequently used on the 2020 campaign trail was mocking his Democratic, Joe Biden, for supposedly misstating the name of whatever location he was speaking in, claiming that ...

Figure 1: These are examples for all the 5 categories. Above each part are claim text and claim image, and below are corresponding document image and document text. There are 5 categories in the dataset, and the categories are divided according to the similarity between the text and image of the claim and the corresponding document image and image.

categories: (1) The first one is traditional learning methods [4, 5, 6, 7], which design plenty of hand-crafted features from the media content of posts and the social context of users. With these sophisticated features, SVM classifiers [4, 7] and decision tree [5, 6] have been trained to debunk fake news. However, the content of fake news is highly complicated and hard to be fully captured by hand-crafted features. (2)With deep neural networks having yielded immense success in learning image and textual representations and their downstream tasks [8, 9], researchers realize that deep learning plays a very important role in detecting fake news. Thus the deep learning based methods [7, 10, 11] are proposed to automatically capture the deep features in an end-to-end way. For example, Ma et al. [7] employ Recurrent Neural Networks (RNNs) to learn the hidden features from posts. Yu et al. [11] use Convolutional Neural Networks (CNNs) to obtain key features and their high-level interactions from fake news. However, most of the above methods focus only on textual content and ignore posts with multi-modal information (such as text, images, etc.), which is a key component of social media platforms.

De-Factify2 [12] is a competition hosted by AAAI 2023 workshop on multi-modal fact checking and hate speech detection, an extension of the De-Factify [13] competition. This workshop aims to encourage researchers from inter-disciplinary domains working on multi-modality and/or fact checking to come together and work on multi-modal (images, memes, videos) fact checking. The goal of this competition is to design a method to classify the given text and images into one of the five categories: Support_Multimodal, Support_Text, Insufficient_Multimodal, Insufficient_Text, and Refute, as displayed in Figure 1. For more details, we refer readers to [12]. To tackle the problem, this paper proposes a Multimodal Attention and Fusion Network (MAFN) with pre-trained models and co-attention networks to perform the shared task, which first extracts features from both text and images, then fuses this information through the co-attention module. Specifically, two powerful Transformer-based pre-trained models, DeBERTa [14] and DeiT [15], are adopted to extract features of images and text both from claims and documents, respectively. Based on that, several co-attention modules are designed to fuse the contexts of text and images. Afterwards, we apply self-attention mechanism to get corresponding representative embeddings. Finally, these embeddings are sequentially concatenated to obtain the final embedding to classify the categories of news.

The main results of this paper can be summarized as follows:

- We leverage an ensemble strategy based on different pre-trained models to obtain better representations for the claims and documents.
- We design a multi-modal attention mechanism and a fusion module to learn the semantic correlation at intra-modality (text or images from claims and documents) and the inter-modality dependencies.
- Our ensemble model outperforms the baseline by 17.0% in terms of testing score, while it still has about 7.6% gap compared to the first prize. Besides, a series of ablation studies were further conducted to study the impact of the designed modules on the overall performance of the model.

2. Related Works

2.1. Fake News Detection

Recently, fake news detection with multi-modality has received considerable attentions. Several approaches[16, 17, 18, 19] conduct fake news detection based on the multimedia content and obtain superior performance. Jin et al. [16] propose a multi-modality based fake news detection model, which extracts the multi-modality information including visual, textual and social context features, and then fuses them by attention mechanism. Khattar et al. [17] introduce a multimodal variational autoencoder that learns a shared representation of text and images. Shivangi et al. [18] make use of the pre-trained BERT to learn text features and apply VGG-19 pre-trained on ImageNet dataset to learn image features. Wang et al. [19] design a novel knowledge-driven multimodal graph convolutional network to jointly model the textual information, knowledge concepts and visual information into a unified framework for fake news detection. MCAN [20] adopts a large-scale pre-trained NLP model and a pre-trained computer vision (CV) model to obtain features from text and images, and then fuses them and frequency domain features from images with multiple co-attention layers.

These methods demonstrate that multi-modal content can also help the model to detect fake news. Thus, we design a multimodal attention and fusion network to mine the semantic correlation among multimedia to facilitate the fact verification.

2.2. Large-Scale Pre-trained Models

Pre-trained models have achieved significant success across numerous tasks. Transformer [21] first introduced in machine translation, has inspired many competitive approaches in natural language processing (NLP) and computer vision tasks. Specifically, Transformer-based pre-trained language models (PLMs) have significantly improved the performance of various NLP tasks due to the ability to understand contextualized information from the pre-trained dataset. GPT [22] replaces bi-LSTMs with a left-to-right Transformer to better extract contextual semantics by a global attention mechanism. DeBERTa [14] proposes a novel disentangled attention mechanism and a new virtual adversarial training to significantly improve the efficiency of pre-training and the performance of 2 downstream tasks.

Vision Transformer (ViT) [23] is a Transformer encoder architecture with patching raw images to achieve competitive results of image classification, compared to state-of-the-art convolutional networks, which demonstrates that convolution-free networks can still capture the visual relation effectively. Then several follow-up studies based on ViT have been conducted. For example, DeiT [15] develops a novel distillation procedure to ensure the student learns better knowledge from the teacher through attention.

In a word, pre-trained models can benefit the procedure of capturing rich information for downstream tasks and also reduce the cost of training from scratch. These advantages drives us to obtain better contextual embedding of images and text with recent pre-trained models.

2.3. Attention Mechanism

Attention mechanisms are demonstrated effective in various tasks such as image captioning [24], machine translation [25] and recommendation system [26]. Concretely, Bahdanau et al. [25] firstly introduce attention in the machine translation task to allow the model to automatically search for parts of a source sentence that are relevant to predicting a target word. Recently, attention mechanisms have been incorporated into fake news detection. For example, Chen et al. [27] propose a deep attention model on the basis of recurrent neural networks (RNN) to learn selectively temporal hidden representations of sequential posts for identifying fake news.

Inspired by the successful applications of attention mechanism, we introduce a co-attention network to compute the intra-modality relationship and inter-modality relationship of image tokens and text words.

3. Method

3.1. Overview

Let $X = \{CT_i, CI_i, DT_i, DI_i\}_{i=1}^{N}$ denote a set of N training data, where the i-th sample is composed of the claim text CT_i, the claim image CI_i, the document text DT_i, and the document image DI_i. $Y = \{y_1, y_2, \cdots, y_N\}_{i=1}^{N}$ denote a set of corresponding labels where $y_i \in \{Support\_Multimodal, Support\_Text, Insufficient\_Multimodal, Insufficient\_Text, Refute\}$. The task of this competition is to classify the data sample into one of the five categories when given a textual claim, claim image, document text and document image.

3.2. Overall Framework

Inspired by [28], we introduce a Multimodal Attention and Fusion Network (MAFN) to improve the performance of multimodal fact verification. By exploiting a multi-modal attention network for multi-modal feature fusion, our model can capture the intra-modality and inter-modality

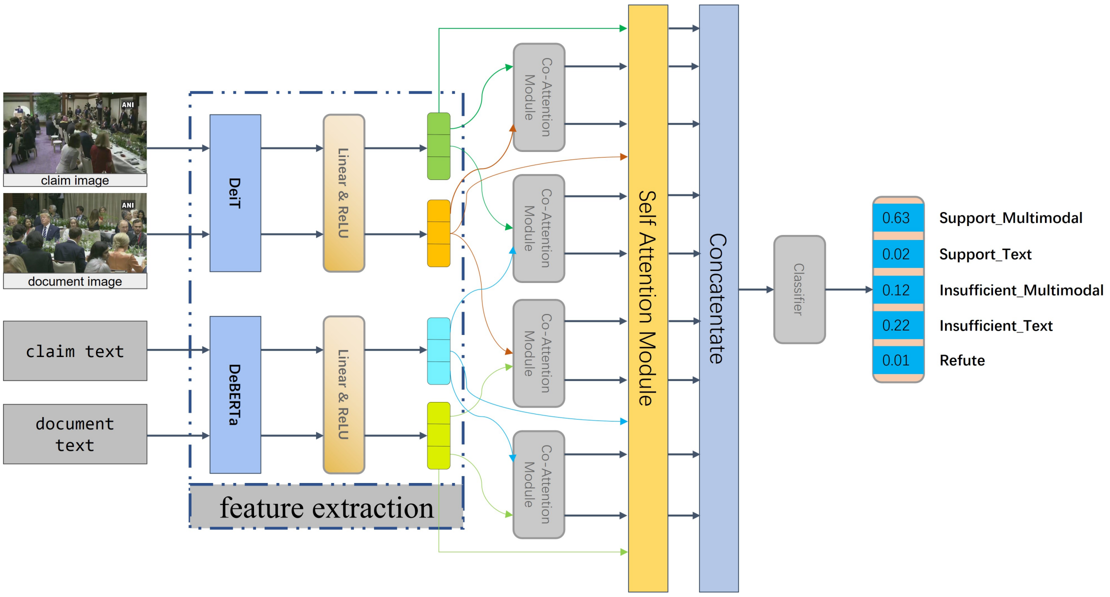


Figure 2: Illustration of the MAFN framework. The feature extraction part aims to transform text and images into corresponding embeddings. The Co-Attention module fuses this information from the same modality (images/text from the claim and document) and different modalities (images and text from the claim/document) to obtain contexts. The self-attention module was used to determine which tokens in the sequence are important and then obtain representative features. Finally, those representative features are concatenated together to predict the possible categories via a category classifier.

relationship of textual and visual content of fake news. The overall architecture is illustrated in Figure 2. Specifically, our model consists of the following components:

- **Text and Image Encoding Network:** The enrichment of pre-trained models enables us to extract rich information without training from scratch. We first use DeBERTa [14] as our pre-trained NLP model and DeiT [15] as our pre-trained CV model to precisely capture the semantics both from the text and the image, and then employ a full connection layer followed by a ReLU function to further extract the multi-modal embedding.
- **Multi-Modality Fusion Network:** As the intra-modality (images/text from the claim and document) or inter-modality (images and text from the claim/document) relationships can facilitate the detection of fake news, we use the multi-modality fusion part to fuse the information from the same modality and different modalities.
- **Category Classifier** aims to classify each piece of data in the dataset into one of five categories with a fully-connected layer followed by a corresponding activation function.

3.3. Text and Image Encoding Network

Text Encoding Network: In order to represent the rich semantic information of sentences, we employ DeBERTa as the core module of our textual language model. Given a sentence, we split it into L words with tokenization technique $T = \{t_1, t_2, \cdots, t_L\}$, and we denote the transformed feature as $S = \{s_1, \cdots, s_L\}$ with s_i corresponding to the transformed feature of t_i. The word representation s_i is calculated by DeBERTa:

$$S = \{s_1, \cdots, s_L\} = DeBERTa(T), \tag{1}$$

where $s_i \in \mathbb{R}^{d_w}$ is the last hidden state of corresponding token in DeBERTa, and d_w is the dimension of the word embedding. Specifically, we feed the claim text and document text into DeBERTa respectively, the corresponding features, e.g. $S_{CT} = DeBERTa(T_{CT})$, $S_{DT} = DeBERTa(T_{DT})$, where the output dimensions of DeBERTa is 768. Then we use the embedding layer for transforming pre-trained embeddings to embeddings in our task. Sepecifically, output of the embedding layer is calculated as follows:

$$\begin{aligned} E_{CT} &= Emb\ (S_{CT}), \\ E_{DT} &= Emb\ (S_{DT}). \end{aligned} \tag{2}$$

Here the Emb is composed of a fully-connected layer and an activation function, and E_{CT}, E_{DT} are d dimension vectors. It is noted that the activation functions in Emb we used are ReLU and Mish [29] for testing the results.

Image Encoding Network: For each input of image, we use pre-trained DeiT model to extract token features. The output is a set of token features $O = \{o_1, \cdots, o_m\}$, where m denotes the token number of the image. The parameters of the pre-trained DeiT are frozen, which means we do not update the parameters of the pretrained model during training. In other words, given the image I, the operation of feature extraction can be expressed as:

$$O = \{o_1, \cdots, o_m\} = DeiT(I), \tag{3}$$

where $o_i \in \mathbb{R}^{d_r}$ and d_r is the dimension of the image embedding. Specifically, we feed the claim image and document image into DeiT respectively, and get the corresponding features, e.g. $O_{CI} = DeiT(I_{CI})$, $O_{DI} = DeiT(I_{DI})$, where the output dimensions of DeiT is 768. Then we use the embedding layer for transforming pre-trained embeddings to embeddings in our task. Sepecifically, output of the embedding layer is calculated as follows:

$$\begin{aligned} E_{CI} &= Emb\ (O_{CI}), \\ E_{DI} &= Emb\ (O_{DI}), \end{aligned} \tag{4}$$

where Emb module is same as the Emb in equation 2. E_{CI}, E_{DI} are d dimension vectors.

3.4. Multi-Modality Fusion

Co-attention block has been widely used in VQA tasks [30], as it can capture dependencies of different inputs. Thus, after generating embeddings of text and images, we adopt multiple co-attention layers as [20, 31] to fuse the embeddings for the improvement of the intra- /inter-modality relations on the detection of fake news.

First, we employ a co-attention layer to separately fuse 1) images of claims and images of documents and 2) text of claims and text of documents(fuse features from same modality). Then we learn the inter-modal alignment by fusing features from different modalities (images and text from the claim/document). Besides, the relation between text and images from the claims or document can be viewed as checking whether they are relative or not. Therefore, we also adopt the co-attention layer for fusing 3) images and text of claims and 4) images and text of documents(fuse features from different modality).

Therefore, we use the co-attention layer for fusing. Specifically, each co-attention layer takes two inputs E_A and E_B to produce two outputs H_A, H_B. We first project E_A/E_B into query $Q \in \mathbb{R}^{N\times d}$, key $K \in \mathbb{R}^{N\times d}$ and value $V \in \mathbb{R}^{N\times d}$ matrices:

$$\begin{aligned} Q_A = E_A W^{Q_A}, K_A = E_A W^{K_A}, V_A = E_A W^{V_A}, \\ Q_B = E_B W^{Q_B}, K_B = E_B W^{K_B}, V_B = E_B W^{V_B}, \end{aligned} \tag{5}$$

where $W^{Q_A}, W^{K_A}, W^{V_A}, W^{Q_B}, W^{K_B}, W^{V_B} \in \mathbb{R}^{d\times d}$.

We then employ attention mechanism together with the residual connection to provide additional capacity for more complex reasoning in our aggregation functions. The specific expression is:

$$\begin{aligned} \tilde{H_A} = LN(E_A + softmax(\frac{Q_A K_B^T}{\sqrt{d}})V_B), \\ \tilde{H_B} = LN(E_B + softmax(\frac{Q_B K_A^T}{\sqrt{d}})V_A), \end{aligned} \tag{6}$$

$$\begin{aligned} H_A = LN(\tilde{H_A} + FFN(\tilde{H_A})), \\ H_B = LN(\tilde{H_B} + FFN(\tilde{H_B})), \end{aligned} \tag{7}$$

where LN is a Layer Normalization and FFN is the same feed forward network as [21]. Now we can use co-attention layer to fuse features from same modalities (or different modalities):

$$\begin{aligned} H_{CIDI}, H_{DICI} = CoAtt(E_{CI}, E_{DI}), \\ H_{CTDT}, H_{DTCT} = CoAtt(E_{CT}, E_{DT}), \end{aligned} \tag{8}$$

$$\begin{aligned} H_{CICT}, H_{CTCI} = CoAtt(E_{CI}, E_{CT}), \\ H_{DIDT}, H_{DTDI} = CoAtt(E_{DI}, E_{DT}), \end{aligned} \tag{9}$$

where $CoAtt$ denotes the co-attention layer.

Afterwards, the aggregation function is adopted to aggregate fused tokens into a representative token. That is, given a fused embedding with $f = \{f_1, \cdots, f_N\} \in \mathbb{R}^{N\times d}$, where N is the sequence length, we perform self-attention mechanism [21] over the fused tokens, which adopts average feature $\overline{\boldsymbol{f}} = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{f}_i$ as the query and aggregates all the tokens to obtain a representative token. Besides, we also feed $E_{CI}, E_{CT}, E_{DI}, E_{DT}$ into the aggregation function for classification.

3.5. Category Classifier

As The features fused by the co-attention layer can represent the complex relationship between claim and document, we first concatenate 8 aggregated outputs H_{CIDI}, H_{DICI}, H_{CTDT}, H_{DTCT}, H_{CICT}, H_{CTCI}, H_{DTDI}, H_{DIDT} from the co-attention layers $H_f = Concat(H_{CIDI} :$

$H_{DICI} : H_{CTDT} : H_{DTCT} : H_{CICT} : H_{CTCI} : H_{DTDI} : H_{DIDT})$. It is worth noting that we also use the outputs of aggregated embeddings since the original information can provide some clues for classifying the news, thus we concatenate 4 aggregated embeddings $E_f = Concat(E_{CI} : E_{CT} : E_{DI} : E_{DT})$. Then we concatenate these two features $Z = Concat(H_f : E_f)$ and feed Z to the subsequent category classification network to predict the label of the given claims and documents. Afterwards, the output of the classifier is the probability as follows:

$$\begin{aligned} Z^{(1)} &= \sigma(ZW^{(0)}), \\ Z^{(2)} &= \sigma(Z^{(1)}W^{(1)}), \end{aligned} \tag{10}$$

$$\hat{y} = softmax(Z^{(2)}W^{(2)}), \tag{11}$$

where $W^{(0)} \in \mathbb{R}^{12d \times d}$, $W^{(1)} \in \mathbb{R}^{d \times d_1}$, and $W^{(2)} \in \mathbb{R}^{d_1 \times 5}$. Note that σ is the same as in Emb, which uses both ReLU and Mish for testing the results.

In the end, We minimize cross-entropy loss $\mathcal{L}$ to verify a multimodal claim:

$$\mathcal{L} = -\sum_{i=1}^{|B|} y_i log(\hat{y}_i). \tag{12}$$

3.6. Ensemble Method

Each classifier may have its strengths and weakness, and ensemble methods have been widely used to enhance the performance. Some models have a higher score on the validation set, we naturally want it to have a larger weight in the final integrated model, thus we use different weights to integrate the model. The formula is derived as follows:

$$p = p_1 \times w_1 + p_2 \times w_2 + \cdots + p_k \times w_k, \tag{13}$$

where $p_1, \cdots, p_k$ are the predicted probability from the corresponding model, $w_1, \cdots, w_k$ are weights with respect to the corresponding model, k is the number of trained models. It is noted that the weight parameters are tuned by hand.

4. Experiments

4.1. Dataset and Implementation

Dataset. Factify [12, 32] is a dataset for multi-modal fact verification, which contains images of the claim, textual claims, reference textual documents and images. Each data contains a reliable source of information, called a “document” and another source whose validity must be assessed, called a “claim”. Both source and claim information sources have a corresponding image. Each data sample belongs to one of the five categories, which are Support_Text, Support_Multimodal, Insufficient_Text, Insufficient_Multimodal and Refute. The labels are defined as:

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

- Insufficient_Multimodal: the claim text is neither supported nor refuted by the document but images are similar to the document.
- Insufficient_Text: both text and images of the claim are neither supported nor refuted by the document, although it is possible that the text claim has common words with the document text.
- Refute: the images and/or text from the claim and document are completely contradictory i.e, the claim is false/fake.

The training set contains 35,000 samples with 5,000 samples per class, and the validation set includes 7,500 samples with 1,500 samples per class. The test set, which is used to evaluate the private score, also contains 7,500 samples. For more details, we refer readers to [12, 33].

Implementation Details. The dimension d was set to 512, the hidden dim of the fully connected layer was set to 1024, the output dimension of DeBERTa and DeiT was 768, and the number of heads was set to 4. The dropout rate was 0.1, and the max sequence length was 512. The batch size was 64, the learning rates were set to 2e-5, the number of training epochs was 30, and the seeds were tested with 24. The weight coefficients between different models are set to 0.7, 0.5, 0.6, 0.7, 0.6, which were manually tuned by validation score. The pre-trained DeBERTa was deberta-base[1], and the DeiT was deit-base-patch16-224[2]. The parameters of the two pre-trained models were frozen during training, which means we do not update their parameters during training. All images were transformed by resizing to 256, center cropping to 224, and normalizing. We preprocessed only for transforming images, and then we stored the text and processed images in corresponding pickle files for training and evaluating. All expriments were conducted with Nvidia GeForce RTX A6000.

Evaluation Metric. The weighted average F1 score across 5 categories is adopted to evaluate the performance.

4.2. Testing Performance

Table 1 shows the performance of the testing set. Our approach achieved 76.051% of the F1-score, winning the fifth prize in detecting fake news. This result outperformed the baseline by 17.0%, but it still has about 7.6% gap compared to the first prize. We think it may be because the pre-trained model is not powerful enough, or the data pre-processing is not enough. Despite the above disadvantages, our approach still demonstrates that using only text and images can achieve competitive performance.

4.3. Ablation Study

To study the impact of each module, we carried a series of ablation studies to verify the effectiveness of the designed modules. As shown in Table 3, applying co-attention only on the same modality (w/o CoAtt(A, B)) is insufficient, which demonstrates the need for modeling dependencies between different modalities. In addition, if only apply co-attention on the different modality (w/o CoAtt(A, A)), the model will not be able to distinguish the difference

[1] https://huggingface.co/microsoft/deberta-base

[2] https://huggingface.co/facebook/deit-base-patch16-224

Table 1
Performance of our model in terms of testing score. Our method achieved fifth prize and we outperformed the baseline by 17.6%.

Rank	Team	Support Text (%)	Support Multimodal (%)	Insufficient Text (%)	Insufficient Multimodal (%)	Refute (%)	Final (%)
1	Triple-Check	82.767	91.383	85.189	89.217	100.00	81.820
2	INO	81.235	90.029	88.807	85.233	99.933	80.795
3	Logically	80.383	90.511	84.393	85.627	98.512	78.967
4	zhang	76.645	87.850	81.610	87.934	99.933	77.423
5	**gzw**	**78.493**	**86.321**	**81.423**	**83.268**	**100.00**	**76.051**
-	Baseline	50.000	82.721	80.240	75.931	98.820	64.990

Table 2
The performance of the five models on the validation set, Ensemble represents using Eq.13 to ensemble model.

Model	Weighted F1(%)	Ensemble
model1	74.048	
model2	70.731	
model3	73.008	76.642
model4	74.937	
model5	72.867	

between claim and document, which will also affect performance. Finally, if removing the co-attention module completely (w/o CoAtt), the performance will drop drastically, which justifies the use of co-attention on the same modality and different modality.

We also explored the effectiveness of the self-attention module. If it is replaced by a simple mean operation, a large performance drop can be observed (see in Table 4), which proves that the model can focus on important sequences through the self-attention module. Meanwhile, it is evident that without concatenating E_f to the final embedding Z, the performance will obviously degrades.

It is noted that our ensemble method slightly improves the performance compared to Pre-CoFact. Our ensemble method includes MAFN (model1 in Table 2), MAFN with replacing DeBERTa with XLM-RoBERTa (model2 in Table 2), MAFN with replacing DeBERTa with RoBERTa (model3 in Table 2), MAFN with replacing DeBERTa with RoBERTa and replacing ReLU with Mish (model4 in Table 2), and MAFN with replacing ReLU with Mish (model4 in Table 2). We the performance of each model in Table 2, and we ensemble the model using equation (13).

4.4. Visualization

Figure 3 visualizes some verification examples by our model and the baseline. It can be observed that our model is superior to the baseline on the multimodal fact verification. On the left side of Figure 3, we can intuitively see that the content of the two pictures is similar, but for the text,

Table 3

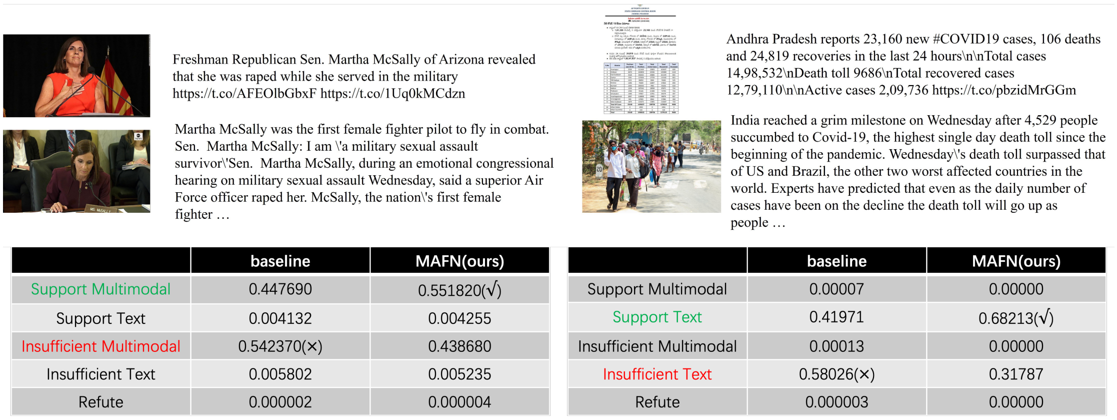
Ablation study of our model in terms of validation score. w/o CoAtt denotes for not using the Co-Attention module, w/o CoAtt(A, B) denotes using only the same modality (Equ. 8) and w/o CoAtt(A, A) denotes using only the different modality (Equ. 9)

Model	w/o CoAtt	w/o CoAtt(A, B)	w/o CoATT(A, A)	MAFN (Ours)
Weighted F1 (%)	72.26 (-4.34)	72.93 (-2.01)	73.32 (-1.62)	74.94

Table 4

Ablation study of our model in terms of validation score. w/o concat E_f denotes not to concatenate E_f to the final feature Z, Mean denotes using mean aggregation to obtain a representative token.

Model	w/o concat E_f	Mean	MAFN (Ours)
Weighted F1 (%)	74.13 (-0.81)	72.87 (-2.07)	74.94



Freshman Republican Sen. Martha McSally of Arizona revealed that she was raped while she served in the military https://t.co/AFEOlbGbxF https://t.co/1Uq0kMCdzn

Martha McSally was the first female fighter pilot to fly in combat. Sen. Martha McSally: I am \'a military sexual assault survivor\'Sen. Martha McSally, during an emotional congressional hearing on military sexual assault Wednesday, said a superior Air Force officer raped her. McSally, the nation\'s first female fighter …

	baseline	MAFN(ours)
Support Multimodal	0.447690	0.551820(√)
Support Text	0.004132	0.004255
Insufficient Multimodal	0.542370(×)	0.438680
Insufficient Text	0.005802	0.005235
Refute	0.000002	0.000004

Andhra Pradesh reports 23,160 new #COVID19 cases, 106 deaths and 24,819 recoveries in the last 24 hours\n\nTotal cases 14,98,532\nDeath toll 9686\nTotal recovered cases 12,79,110\n\nActive cases 2,09,736 https://t.co/pbzidMrGGm

India reached a grim milestone on Wednesday after 4,529 people succumbed to Covid-19, the highest single day death toll since the beginning of the pandemic. Wednesday\'s death toll surpassed that of US and Brazil, the other two worst affected countries in the world. Experts have predicted that even as the daily number of cases have been on the decline the death toll will go up as people …

	baseline	MAFN(ours)
Support Multimodal	0.00007	0.00000
Support Text	0.41971	0.68213(√)
Insufficient Multimodal	0.00013	0.00000
Insufficient Text	0.58026(×)	0.31787
Refute	0.000003	0.00000

Figure 3: Some examples of classification results of our model and baseline model. The top is the claim image and text, the middle is the document image and text, and the bottom is the five categories of probabilities output by our model and the baseline model, where the green font represents the ground truth label, and the red font represents the label misjudged by the baseline model.

the claim and document are different in length, and the sentence structure is also very different, but the semantics are the same. Our model can correctly classify the results, demonstrating that our model can learn high-level semantic connections between claim and document texts, which we attribute to the use of Co-Attention module. The example on the right also shows that our model can understand high-level semantic information.

5. Conclusion

In this paper, we proposed a multimodal fact verification method called MAFN, which utilizes pre-trained models and multiple co-attention networks to alleviate the effect of fake news. To further improve the performance, we adopted an ensemble method by weighting several different pretrained models. The ablation study demonstrates the effectiveness of our proposed approach. The test scores can also illustrates the effectiveness of our model.

References

[1] W. Zheng, G. Zhenwei, X. Xing, L. Yadan, Y. Yang, S. Heng Tao, Point to rectangle matching for image text retrieval, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, p. 4977–4986.

[2] W. Zheng, Z. Jie, M. Jing, L. Jingjing, A. Jiangbo, Y. Yang, Discovering attractive segments in the user-generated video streams, Information Processing & Management 57 (2020) 102–130.

[3] W. Zheng, Y. Yang, L. Jingjing, X. Zhu, Universal adversarial perturbations generative network, World Wide Web 25 (2022) 1725–1746.

[4] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th international conference on World wide web, 2011, pp. 675–684.

[5] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent features of rumor propagation in online social media, in: 2013 IEEE 13th international conference on data mining, IEEE, 2013, pp. 1103–1108.

[6] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, S. Shah, Real-time rumor debunking on twitter, in: Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 1867–1870.

[7] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks (2016).

[8] J. Wei, Y. Yang, X. Xu, X. Zhu, H. T. Shen, Universal weighting metric learning for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2022) 6534–6545. doi:10.1109/TPAMI.2021.3088863.

[9] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, H. T. Shen, Universal weighting metric learning for cross-modal matching, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13005–13014.

[10] J. Ma, W. Gao, K.-F. Wong, Detect rumors on twitter by promoting information campaigns with generative adversarial learning, in: The world wide Web conference, 2019, pp. 3049–3055.

[11] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, et al., A convolutional approach for misinformation identification., in: IJCAI, 2017, pp. 3901–3907.

[12] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. R. Anku Rani, 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.

[13] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Benchmarking multi-modal entailment for fact verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, ceur, 2022.

[14] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: 9th International Conference on Learning Representations, 2021.

[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, 2021, pp. 10347–10357.

[16] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 795–816.

[17] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The world wide web conference, 2019, pp. 2915–2921.

[18] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, P. Kumaraguru, Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract), in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13915–13916.

[19] Y. Wang, S. Qian, J. Hu, Q. Fang, C. Xu, Fake news detection via knowledge-driven multimodal graph convolutional networks, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 540–547.

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

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 2017, pp. 5998–6008.

[22] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[23] 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: 9th International Conference on Learning Representations, 2021.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.

[25] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[26] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, T.-S. Chua, Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention, in: Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, 2017, pp. 335–344.

[27] T. Chen, X. Li, H. Yin, J. Zhang, Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, in: Pacific-Asia conference on knowledge discovery and data mining, Springer, 2018, pp. 40–52.

[28] W.-Y. Wang, W.-C. Peng, Team yao at factify 2022: Utilizing pre-trained models and

co-attention networks for multi-modal fact verification, arXiv preprint arXiv:2201.11664 (2022).

[29] D. Misra, Mish: A self regularized non-monotonic neural activation function, CoRR abs/1908.08681 (2019).

[30] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, H. Li, Dynamic fusion with intra- and inter-modality attention flow for visual question answering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6639–6648.

[31] N. Wang, Z. Wang, X. Xu, F. Shen, Y. Yang, H. T. Shen, Attention-based relation reasoning network for video-text retrieval, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1–6. doi:10.1109/ICME51207.2021.9428215.

[32] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Factify: A multi-modal fact verification dataset (2022).

[33] 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.