Graph Interactive Network with Adaptive Gradient for Multi-Modal Rumor Detection

Tiening Sun
Soochow University
Suzhou, China
tnsun@stu.suda.edu.cn

Zhong Qian
Soochow University
Suzhou, China
qianzhong@suda.edu.cn

Peifeng Li*
Soochow University
Suzhou, China
pfli@suda.edu.cn

Qiaoming Zhu
Soochow University
Suzhou, China
qmzhu@suda.edu.cn

ABSTRACT

With more and more messages in the form of text and image being spread on the Internet, multi-modal rumor detection has become the focus of recent research. However, most of the existing methods simply concatenate or fuse image features with text features, which can not fully explore the interaction between modalities. Meanwhile, they ignore the convergence inconsistency problem between strong and weak modalities, that is, the dominant rumor text modality may inhibit the optimization of image modality. In this paper, we investigate multi-modal rumor detection from a novel perspective, and propose a Multi-modal Graph Interactive Network with Adaptive Gradient (MGIN-AG) to solve the problem of insufficient information mining within and between modalities, and alleviate the optimization imbalance. Specifically, we first construct fine-grained graph for each rumor text or image to explicitly capture the relation between text tokens or image patches in unimodal. Then, the cross modal interaction graph between text and image is designed to implicitly mine the text-image interaction, especially focusing on the consistency and mutual enhancement between image patches and text tokens. Furthermore, we extract the embedded text in images as an important supplement to improve the performance of the model. Finally, a strategy of dynamically adjusting the model gradient is introduced to alleviate the under optimization problem of weak modalities in the multi-modal rumor detection task. Extensive experiments demonstrate the superiority of our model in comparison with the state-of-the-art baselines.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; *Machine learning*; • **Information systems** → *World Wide Web*.

*Corresponding Author.

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592250

KEYWORDS

rumor detection, fake news detection, multi-modal fusion, social networks, graph neural networks

ACM Reference Format:
Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2023. Graph Interactive Network with Adaptive Gradient for Multi-Modal Rumor Detection. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592250

1 INTRODUCTION

Since human society entered the Internet era, profound changes have taken place in the information dissemination mechanism. Especially with the rapid development of multimedia social platforms (e.g., Weibo and Twitter), everyone can be the producer, communicator and receiver of information, which provides convenient conditions for the breeding and spreading of rumors. Rumors are good at leading the public's attention away from facts to disinformation. Meanwhile, due to their fast spreading characteristics, they can easily reach a scale sufficient to dominate public discourse and change public opinion, which can cause a continuous negative impact on society. For example, panic selling was triggered in the US stock market due to fake news about President Obama's injury in 2013 [11]. The result of the US presidential election were affected as public opinion was shaken by virtual bots spreading political rumors in 2016 [1]. Hence, in the context of the new media era, how to quickly and effectively identify online rumors has become a research hotspot.

In the early years, text content was the main manifestation of rumors, so conventional rumor detection methods mainly focused on mining rumor text content and used advanced deep learning technology [14, 20, 34] to obtain high-dimensional feature encoding. However, with the development of multimedia technology, rumor posts have evolved from a single plain text to a multi-modal form consisting of text, images and even videos, which makes the conventional text-oriented models are not competent for the task of multi-modal rumor detection. Fortunately, the research on rumor detection using multi-modal information has also made preliminary progress in recent years. EANN [28] fuses the features of text and images via simple concatenation operation, and introduces adversarial learning to enhance feature representation. MCAN [32]

(a) token-patch relationship (b) rumor image with additional text

Figure 1: Two multi-modal rumor instances. (a) The framed patches in the image and the tokens marked in the claim are regarded as important cues with strong associations. (b) The rumor image carries important textual information.

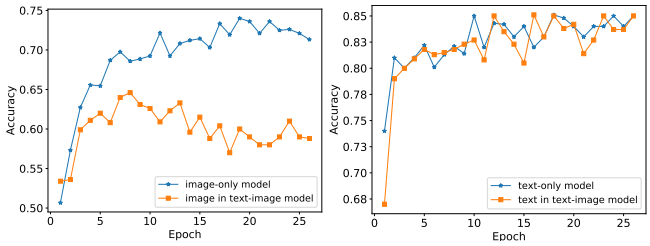

Figure 2: The performance of rumor images and claim texts from the PHEME dataset in the uni-modal model and fused-modal model, respectively. The text modality, which is dominant in rumor detection, is not disturbed too much in performance, but the performance of the weak image modality is degraded after concatenation fusion operation.

extracts spatial-domain and frequency-domain features from images, and then uses an attention mechanism to fuse image features and text features at a coarse-grained level. MFAN [37] constructs a multi-modal heterogeneous graph from a global perspective, and utilizes graph neural network to encode it for rumor classification.

Although the multi-modal features have been paid more attention, the existing methods still have the following limitations: (1) most models simply concatenate text features and image features, or use coarse-grained fusion strategies, which cannot fully mine effective information within and between modalities. More specifically, they often ignore the interactions and long-distance dependencies between text tokens and image patches that may become evidence clues, when judging the authenticity of a message (assumed to consist of only text and images). For example, the "*street property*" token marked in bold in Figure 1(a) not only has a dependency on the "*100 million*" at the end of the sentence, but also has a strong correlation with the "*nuts cake*" patches scattered in the image, which can be used as important clues. Coarse-grained encoding and fusion cannot capture the interaction between image patches or text tokens, and ignore the co-occurrence features or consistencies between patches and tokens, which are regarded as the important multi-modal cues to correctly classify rumors. (2) Generally, text is the main manifestation of rumors, while images are mostly auxiliary, which may lead to the rumor detector being dominated by text features in the training process and inhibit the optimization of image features, as shown in Figure 2. Hence, it is important to dynamically adjust the convergence rate of different modalities to help prevent weak modality from falling into under-optimization. (3) The text embedded in the rumor image in Figure 1(b) has proved to be one of the key clues, but how to more effectively fuse the feature into the rumor detector remains to be investigated.

To address the above challenges, we propose a Multi-modal Graph Interactive Network with Adaptive Gradient (MGIN-AG) for multi-modal rumor detection. Specifically, the MGIN-AG framework is mainly composed of three design strategies as follows. (1) To capture the interaction relations within modalities, we first construct a graph for each text based on the dependency tree to mine the dependencies between words. Then, to obtain the relations between image patches, we construct a fully connected graph, in which the patches are nodes and the edge weights are determined by the similarity between patches. Furthermore, to learn the long-distance dependencies between modalities, we explicitly connect the above image patches with text tokens to construct a cross-modality heterogeneous graph, and design an attention aggregation layer to calculate and analyze the implicit consistency between modalities. Finally, the text modality graph, image modality graph and cross-modality graph are fed to the Graph Network [10] to obtain three different hierarchical graph representations. (2) The embedded text in the image is one of the important clues, which is extracted by OCR technology. We use the self-attention mechanism to learn the interaction between words in the embedded text, and also design a claim-guided co-attention mechanism to choose which words are more important for rumor classification. (3) To ensure that the features of different modalities in MGIN-AG can be well optimized, we extend the dual-modal mechanism in OGM [17] into a multi-modal mechanism, named MOE, which can adaptively adjust the corresponding gradients according to the convergence rate between different modules of the model. Our contributions are summarized as follows.

- We propose a fine-grained multi-modal graph interaction network for multimedia rumor detection, which can not only explicitly learn the dependencies between text tokens and image patches from the graph perspective, but also implicitly mine interactions between different modalities, especially the consistency and mutual enhancement.
- We introduce a gradient adjustment strategy to balance the optimization process of different modalities, while the embedded text in images is more effectively fused into the rumor detector, thereby improving the model competitiveness.
- We experimentally demonstrate that our model outperforms state-of-the-art baselines on real-world datasets.

2 BACKGROUND

2.1 Problem Definition

Multi-modal rumor detection is defined as a classification problem, whose purpose is to learn a classifier from a set of labeled training

posts (where each post consists of an image and a text), and then use it to predict the label of the test posts in this paper. Specifically, given a multi-media post $P = \{I, C, T\}$ which consists of the image I, claim C and embedded text T extracted from the image I, the goal is to learn a model $f(\cdot)$ to classify the post P into the ground-truth label $y \in \{0, 1\}$, where 0 denotes rumor, and 1 denotes non-rumor.

2.2 Related Work

Most of the popular rumor detection models rely on deep learning technology and focus on extracting and analyzing rumor text features. Many previous studies mainly used neural network encoders such as RNN or CNN directly to generate high-dimensional representations of claims for training and rumor classification [12–14, 16, 31, 34]. With the popularity of pre-trained models, BERT-based methods are becoming mainstream [5, 6, 23]. For example, Dun et al. [5] introduced BERT as the text encoder and used self-attention mechanism to capture the relations between knowledge entities and claim words. In addition, some studies attempted to integrate text information with other rumor clues to improve detection accuracy, such as comment information [23, 31, 33], conversational structure [2, 13, 25, 35] and user stance [15, 18, 30].

With the development of multimedia technology, text-oriented methods obviously cannot be applied to identify the authenticity of multimedia posts composed of text, images or even videos. Hence, Wang et al. [28] proposed a multi-modal rumor detection framework in which image features encoded by VGG-19 are simply concatenated with text features for rumor classification. Khattar et al. [9] added decoders on the basis of [28] to improve the quality of multi-modal representation. However, the direct concatenation operation cannot effectively learn multi-modal mutual interaction and enhancement. To address this problem, cross-modal attention mechanisms are widely exploited. Qian et al. [20] designed a multi-modal contextual attention network that can mine hierarchical semantic relations while modeling multi-modal information. Wu et al. [32] extracted the spatial-domain and frequency-domain features from images and the textual features from text, respectively, which are fused to obtain multi-modal features via multiple co-attention modules. Graph neural networks are also used to aggregate information of different modal nodes. For example, Wang et al. [29] jointly modeled textual information, knowledge concepts and visual information, and adopted simple graph convolution for encoding. Zheng et al. [37] introduced the graphical social context feature to improve model performance.

In addition to the improvement of multi-modal fusion strategies, there are some studies focusing on the inconsistency of image-text. Sun et al. [22] employed the orthogonal constraint to obtain multi-modal shared embedding and unique embedding to measure the inconsistency between modalities. Qi et al. [19] proposed an entity inconsistency framework, in which the similarity between image entities and text entities is used as the cue for classifying rumors.

The uniqueness of our work is that, we constructs fine-grained graphs for images and text, respectively for mining in-modal and cross-modal feature interactions (intuitively, we believe that our approach can implicitly learn multi-modal mutual enhancement and consistency), while an adaptive gradient adjustment strategy is employed to promote feature learning.

3 METHOD

3.1 Model Overview

In this section, we describe in detail our proposed Multi-modal Graph Interactive Network with Adaptive Gradient (MGIN-AG) for multi-modal rumor detection. As shown in Figure 3, the architecture of MGIN-AG mainly consists of five modules: a) *Image-graph representation*, which first constructs a fully connected graph for the image split into multiple patches, and then jointly employs a pre-trained Visual Transformer (ViT) [4] and Graph Convolutional Networks (GCN) [10] to obtain the graph representation. b) *Claim-graph representation*, which first uses Bidirectional Encoder Representations from Transformers (BERT) [3] to encode claim tokens, then constructs a graph for the claim based on the dependencies between words, and finally uses GCN to obtain the graph representation. c) *Cross-modality graph representation*, which constructs a cross-modality graph with image patches and text tokens as nodes, and utilizes the Graph Attention Aggregation (GAA) layer to obtain fused interaction features. d) *Embedded text representation*, which employs two attention mechanisms to capture features. e) *Rumor classification*. A gradient adjustment strategy is introduced to assist weak modules in generating better feature representations.

3.2 Image-Graph Representation

Just as there are grammatical relations between words in text, the mining of associations between image patches should also be paid attention to, as shown in Figure 1(a). Hence we explicitly concatenate image patches and used ViT and GCN to aggregate features. Specifically, given an image I, we first resize the image to 224×224, and split it into $m = k \times k$ patches[1] to obtain a patch sequence $v = \{p_i\}_{i=1}^{m}$. Then the sequence v is fed into the pre-trained ViT model to obtain an image feature matrix $X^V \in \mathbb{R}^{(m+1)\times d^v}$.

$$X^V = \{x_1^v, x_2^v, x_3^v, ..., x_{m+1}^v\} = ViT([class]v), \tag{1}$$

where x^v is the feature vector corresponding to each patch. To ensure the unity of all modal dimensions, a linear transformation is performed on X^V and obtain $V \in \mathbb{R}^{(m+1)\times d^h}$.

$$V = \{v_1, v_2, v_3, ..., v_{m+1}\} = X^V W, \tag{2}$$

where $W \in \mathbb{R}^{d^v \times d^h}$ is a trainable parameter matrix. Note that, the special token $[class]$ will be discarded, and the remaining image patches are utilized to construct the image-modal graph. The representation V is used as the initialization feature of the graph node.

Next, to capture the interactive features between patches, we take image patches as nodes and the similarity between patches as edge weights to construct a homogeneous fully connected graph for each image, and its adjacency matrix $A^V_{\hat{i},\hat{j}}$ is defined as follows.

$$A^V_{\hat{i},\hat{j}} = \begin{cases} 1 & if \ \ \hat{i} = \hat{j} \\ 1 & if \ \ |\hat{i} - \hat{j}| = 1 \ or \ |\hat{i} - \hat{j}| = k \\ sim(\hat{i}, \hat{j}) & otherwise \end{cases}, \tag{3}$$

[1] m is better set to 9 or 16, because the smaller the patch in our work, the less information is covered, which is not conducive to the learning of graph representation.

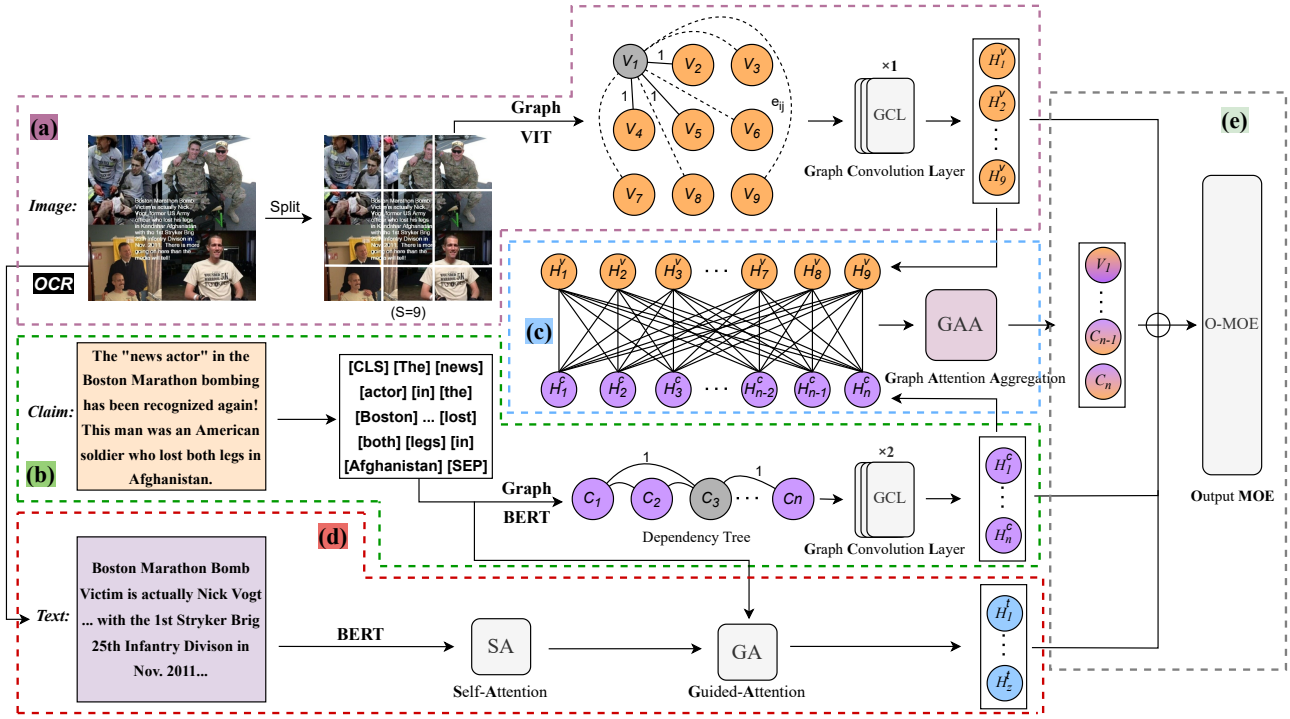

Figure 3: Overview of our MGIN-AG multi-modal rumor detection model. (a) ViT and single-layer GCL are exploited to calculate the image representation. (b) BERT and two-layers GCLs are used to obtain the claim representation. (c) Interaction between the image and the text. (d) Extraction and encoding of embedded text in the image. (e) Classification module, where MOE is the gradient adjustment strategy and ⊕ represents the concatenation operation.

where $|\cdot|$ represents the absolute value calculation, and $sim(\hat{i}, \hat{j})$ is the cosine similarity. Neighboring image patches are coherent, so their edge weights are set to 1, while the edge weights between non-neighboring patches are calculated according to the image similarity. Subsequently, we employ the graph convolution layer (GCL) to aggregate the correlation of nodes, that is, node features V and adjacency matrix $A^V_{\hat{i},\hat{j}}$ are fed to the single-layer[2] GCL to obtain the corresponding graph representation $H^V \in \mathbb{R}^{m \times d^h}$ as follows.

$$H^V = \sigma(\hat{A}^V V W^V), \tag{4}$$

where σ is an activation function such as the ReLU function. $\hat{A}^V = D^{-\frac{1}{2}} A^V_{\hat{i},\hat{j}} D^{-\frac{1}{2}}$ is the normalized adjacency matrix, where D is the degree matrix of $A^V_{\hat{i},\hat{j}}$.

Finally, we use the mean-pooling operator (MEAN) and the skip connection which helps improve training stability to aggregate the information of H^V representing the set of node representations. It is formulated as follows.

$$h^v = MEAN(H^V + V), \tag{5}$$

where $h^v \in \mathbb{R}^{1 \times d^h}$ is the image-graph representation.

[2]The depth of the image-modal graph is 1, so we only use single-layer GCL.

3.3 Claim-Graph Representation

Rumor clues may be expressed in multiple words, so we introduce the dependency tree and jointly employ BERT and GCN to generate augmented features for claim tokens. Specifically, given a claim C containing n words, we first split it into a sequence of words $c = \{w_i\}_{i=1}^n$, then utilize the pre-trained BERT to map each word into a d^c-dimensional embedding as follows.

$$X^C = \{x^c_1, x^c_2, x^c_3, ..., x^c_{n+2}\} = BERT([CLS]c[SEP]). \tag{6}$$

The special tokens [CLS] and [SEP] are discarded, and the remaining word embeddings are fed to a bidirectional LSTM as follows.

$$C = \{c_1, c_2, c_3, ..., c_n\} = LSTM(X^C), \tag{7}$$

where $C \in \mathbb{R}^{n \times d^h}$ is the feature matrix, which is used in constructing the claim graph.

Next, we construct an undirected graph for each claim based on the dependency tree [3], which can effectively capture syntactic and word dependencies in the claim. The adjacency matrix of the claim graph is defined as follows.

$$A^C_{\bar{i},\bar{j}} = \begin{cases} 1 & if \ \ \mathcal{D}(w_{\bar{i}}, w_{\bar{j}}) \\ 1 & if \ \ \bar{i} = \bar{j} \\ 0 & otherwise \end{cases}, \tag{8}$$

[3]In our work, the spaCy toolkit is used to generate dependency trees.

where $\mathcal{D}(w_{\bar{i}}, w_{\bar{j}})$ indicates that w_i and w_j have a relation in the dependency tree. Subsequently, we feed C and $A^C_{\bar{i},\bar{j}}$ into a two-layers GCLs to obtain the corresponding representation $H^C \in \mathbb{R}^{n \times d^h}$ as follows.

$$H^C_{(2)} = \sigma(\hat{A}^C \sigma(\hat{A}^C C W^C_{(0)}) W^C_{(1)}), \tag{9}$$

where σ is ReLU, W^C is the weight matrix, and $\hat{A}^C = D^{-\frac{1}{2}} A^C_{\bar{i},\bar{j}} D^{-\frac{1}{2}}$ is the normalized adjacency matrix, where D is the degree matrix of $A^C_{\bar{i},\bar{j}}$. Then MEAN is used to obtain claim-graph representation $h^c \in \mathbb{R}^{1 \times d^h}$, as follows.

$$h^c = MEAN(H^C_{(2)} + C). \tag{10}$$

3.4 Cross-Modality Graph Representation

The cross-modal interactive relation between images and claims is important, especially to implicitly capture the consistency and mutual enhancement between image patches and claim tokens from a fine-grained perspective, as shown in Figure 1(a). In this section, we construct a cross-modal graph by using image patches and claim tokens as nodes, where each node is only connected to nodes with a different modality from it. For example, for an image patch node p_i, there will be edges between p_i and every claim token w_i, and the p_i will not be connected to the remaining image patch nodes. Note that the self-loop is also not considered. Hence, the adjacency matrix $A^M_{i,j} \in \mathbb{R}^{(m+n)\times(m+n)}$ of the cross-modal heterogeneous graph is defined as follows.

$$A^M_{i,j} = \begin{cases} 1 & if \ \ (i > m, j < m) \ \ or \ \ (i < m, j > m) \\ 0 & otherwise \end{cases}. \tag{11}$$

As shown in Figure 3, we take the features of image patches and claim tokens that have aggregated neighbor information as the initial representation of cross-modal graph nodes. However, due to the node heterogeneity, image nodes and claim nodes have different feature spaces. Therefore, we design a shared linear transformation matrix $M_{share} \in \mathbb{R}^{d^h \times d^h}$ to project the image node feature $h^V_i \in \mathbb{R}^{1 \times d^h}$ from H^V and claim node feature $h^C_i \in \mathbb{R}^{1 \times d^h}$ from H^C into the same feature space as follows.

$$h^{'}_i = h_i \cdot M_{share}, \tag{12}$$

where $h^{'}_i \in \mathbb{R}^{1 \times d^h}$ is the projected feature of node i. Subsequently, we design a Graph Attention Aggregation (GAA) layer to capture the interaction between image patches and claim tokens, where the signed attention [24] is introduced. Formally, we first calculate the positive weight coefficient $a_{i,j}$ and negative weight coefficient $a^{\hat{}}_{i,j}$ between the node i and the node j, respectively, as follows.

$$a_{i,j} = \frac{exp(\sigma(\mu^T \cdot [h^{'}_i \| h^{'}_j]))}{\sum_{u \in \mathcal{N}} exp(\sigma(\mu^T \cdot [h^{'}_i \| h^{'}_u]))}, \tag{13}$$

and

$$\hat{a}_{i,j} = -\frac{exp(-\sigma(\mu^T \cdot [h^{'}_i \| h^{'}_j]))}{\sum_{u \in \mathcal{N}} exp(-\sigma(\mu^T \cdot [h^{'}_i \| h^{'}_u]))}, \tag{14}$$

where $\mathcal{N}$ denotes the neighbors of the node i, and μ is the attention vector. The signed attention is often used in rumor detection based on conversational threads. In our work, it is used to capture the mutual interaction between modalities from multiple perspectives. In other words, the signed attention can use the positive $a_{i,j}$ and the negative $\hat{a}_{i,j}$ respectively to indicate whether the connected the neighbor node j supports or opposes the current node i. Then, the embedding of the node i can be aggregated by the neighbor's features with the corresponding weight coefficients as follows.

$$z_i = \Big\Vert_{h=1}^{H} \sigma([(\sum\nolimits_{j \in \mathcal{N}} a_{i,j} \cdot h^{'}_j) \| (\sum\nolimits_{j \in \mathcal{N}} \hat{a}_{i,j} \cdot h^{'}_j)]), \tag{15}$$

where z_i is the learned feature of the node i, and H is an adjustable hyperparameter. The feature aggregation process of other nodes is similar. Finally, we get the image patch and claim token feature matrix $\hat{H^V} = \{z^v_1, z^v_2, ..., z^v_m\}$ and $\hat{H^C} = \{z^c_1, z^c_2, ..., z^c_n\}$ after interaction, respectively, and use MEAN (similar to Eq. 5) to obtain the cross-modality image and claim representations $\hat{h^v}$ and $\hat{h^c}$.

3.5 Embedded Text Representation

Embedded text on the image has been proved to be one of the important clues to classify rumors. However, how to incorporate its features into the model remains to be studied. In this paper, we employ self-attention (SA) [26] and guided-attention (GA) to better learn the representation of the embedded text. Specifically, We first use OCR [4] technology to extract embedded text from each image. Then, given a piece of embedded text $t = \{w^{'}_i\}^z_{i=1}$, each word $w^{'}_i$ is projected into a word embedding x^t_i via BERT embedding layer. Next, we feed them to a linear layer for dimension unification and obtain a feature matrix $T = \{t_1, t_2, ..., t_z\}$. To learn the interaction between words in the text, we use SA to process the input T as follows.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{16}$$

where $Q = K = V = T$, and d_k is the their dimension. In addition, the formalization of multi-head SA is as follows.

$$MultiHead(Q, K, V) = Concat(Attention_1, ..., Attention_{H^{'}}), \tag{17}$$

where $H^{'}$ is the number of heads. Since not all embedded texts are useful for classification, we propose a claim-guided GA layer that can measure the importance of each word in the embedded text according to the claim. Its calculation process is similar to Eq. 16, but $Q = C$ and $K = V = T$. Finally, we can obtain the representation h^t of the embedded text.

3.6 Rumor Classification

Now, we have obtained the image-graph representation h^v, claim-graph representation h^c, embedded text representation h^t, cross-modality image representation $\hat{h^v}$, and cross-modality claim representation $\hat{h^c}$. Then, we concatenate them to merge the information as follows.

[4] Baidu API is used: https://ai.baidu.com/tech/ocr/

$$h^o = concat(h^c, h^v, h^t, \hat{h^v}, \hat{h^c}). \quad (18)$$

Next, h^o is fed into the full-connection layer and a softmax layer, and the output is calculated as follows.

$$\hat{y} = softmax(W^F h^o + b^F), \quad (19)$$

where $\hat{y}$ is the predicted probability distribution. W^F and b^F are the trainable weight matrix and bias respectively. Finally, we use stochastic gradient descent to minimize the cross-entropy loss for model training.

3.7 MOE Training Strategy

Our proposed framework mainly consists of four modules for learning representations, including image representation, claim representation, embedded text representation, and cross-modal interaction modules. Then, the features generated by these modules are concatenated to obtain a final vector h^o for classification. However, the parameter optimization process in different modules may affect each other as shown in Figure 2. Wang et al. [27] found that different modalities have different convergence speeds during training, and the dominant modality may inhibit the optimization of the weak modality. Peng et al. [17] proposed the OGM method to alleviate the optimization imbalance caused by joint training of dual-modal (vision and audio). It is not feasible to directly introduce the OGM method into our framework, because it only considers two modalities visual and audio, while our modules can be regarded as five modalities in different spaces. Hence, we start with OGM and then design a dynamic gradient adjustment strategy MOE for the multi-modal rumor detection task to balance the optimization process of each module in the proposed framework.

Specifically, we first calculate the contribution score of each module to the optimization objective as follows.

$$score_i^g = \sum_{c=1}^{C} 1_{c=y_i} \cdot softmax(W^g h_i^g + \frac{b}{n})_c, \quad (20)$$

where $g \in \{v, c, t, \hat{v}, \hat{c}\}$ and $score^g$ represents the contribution score of each module. $n = 5$ is the number of representations, and W^g and b are extracted from the weight and bias in the classification layer. C is the number of label categories. y_i is the ground-truth of h_i, and only positive results are summed. Then, we calculate the contribution discrepancy ratio ρ^g of different modules as follow.

$$\rho^g = \frac{\sum_{i \in B} socre_i^g}{\sum_{q \in v,c,t,\hat{v},\hat{c}} \sum_{i \in B} socre_i^q / n}, \quad (21)$$

where B is a random mini-batch. Next, we construct a gradient adjustment factor k^g as follows.

$$k^g = \begin{cases} 1 - tanh(\alpha \cdot \rho^g) & if \ \ \rho^g > 1 \\ 1 & otherwise \end{cases}, \quad (22)$$

where α is a temperature parameter. Finally, following [17], we integrate k^g into the SGD optimization algorithm and add the randomly sampled Gaussian noise. It should be noted that the modules in our framework are all independent, where the module (c) in Figure 3 only uses image and claim graph representations but the gradient back-propagation is not performed.

Table 1: Statistics of the datasets

Statistic	Weibo	PHEME
# source tweets	9528	1198
# rumors	4749	599
# non-rumors	4779	599
# images	9528	1198
# ocr text	9227	827

4 EXPERIMENTATION

In this section, we evaluate our proposed MGIN-AG comparing it with SOTA benchmarks, and give some discussion and analysis. Moreover, we perform ablation analysis to verify the effectiveness of each module of MGIN-AG in turn.

4.1 Datasets

We evaluate MGIN-AG on two public real-world datasets: English PHEME [21] and Chinese Weibo [7], which are collected from the most influential social media sites, Twitter and Weibo, respectively. They all contains only two types of tags: Rumor (R) and Non-Rumor (N), which is used for the binary classification of rumors and non-rumors. We extract the images corresponding to each source text in PHEME from Twitter. To match our proposed framework, only the first image is retained. In addition, we remove samples that do not contain images, while balancing the number of categories. Finally, we use OCR technology to extract the embedded text from PHEME and Weibo images as one of the model input. Table 1 show the statistics of the resulting two datasets after removal.

4.2 Implementation Details

Our MGIN-AG is implemented by PyTorch [8]. For PHEME, because of its small amount of data, we follow [2] and randomly split the dataset into five parts to preform 5-fold cross-validation to obtain more stable and accurate experimental results. In addition, the batch size is set to 32, the learning rate is initialized to 1e-3 and gradually decreases during training according to the decay rate of 1e-4. The α in Eq. 22 is set to 0.2. For Weibo, we split the datasets for training, validation, and testing with a ratio of 6:2:2. Meanwhile, the batch size is set to 8, and the hyperparameter α is set to 0.6. To prevent overfitting, the early stopping strategy is introduced. Then, the SGD is adopted to optimize our objective function. Finally, the Accuracy (Acc.), Precision (Prec.), Recall (Rec.) and F_1-measure (F_1) are used as evaluation metrics in the two datasets.

4.3 Baselines

It's worth noting that our MGIN-AG does not use the reply information, which can act as real-time and early rumor detection. Compared with those models [23, 36, 37] heavily relying on reply information and lacking real-time capability, our MGIN-AG can detect rumor when users post messages and does not need to wait for user's reply information. For fair comparison, we only compare our MGIN-AG with the following state-of-the-art baselines, which do not need reply information, as follow.

- **ViT** [4] is a recently popular visual encoder, and its model architecture is almost exactly the same as Transformer in natural language processing.
- **BERT** [3] is currently the most popular pretrained language representation model.
- **EANN** [28] is a multi-modal rumor classifier, in which VGG-19 and Text-CNN are used to encode visual and text information respectively.
- **MVAE** [9] is a multi-modal variational autoencoder that can effectively learn shared representations between images and text.
- **HMCAN** [20] is the state-of-the-art method, which extracts multi-modal high-order complementary information and hierarchical semantics of text by designing a hierarchical multi-modal contextual attention network.

Among them, ViT and BERT are single-modal methods, where ViT only considers image information, and BERT only considers text information. The other baselines are multi-modal models that consider both images and text.

4.4 Results and Discussion

Tables 2 and 3 show the performance of all baselines on two datasets Weibo and Pheme, where the bold part represents the best performance. We can observe that our MGIN-AG significantly outperforms all the baselines. Unsurprisingly, the ViT model, which only considers visual features, gets the worst results, mainly because rumor detection is a claim-dominated classification task. Hence, the performance of the BERT model with claim information as input is much better than that of ViT. Both EANN and MVAE are simple multi-modal detection frameworks, among which MVAE improves the quality of representation by introducing variational autoencoders, so its performance is better than EANN. However, due to the weak encoder of MVAE, its prediction accuracy is inferior to the pre-trained BERT.

HMCAN is the BERT-based method, which is the state-of-the-art benchmark used to verify the superiority of our MGIN-AG. HMCAN designs a multi-modal contextual attention layer and a contextual transformer to learn the correlation coefficient between images and claims, which show excellent performance.

Our MGIN-AG beats all benchmarks, and its superiority stems from three reasons as follows. 1) The powerful visual encoder ViT and textual encoder BERT can generate higher quality representations. Meanwhile, we construct in-modal graphs for images and claims separately, which can effectively learn the dependencies between image patches or claim tokens from a fine-grained perspective. To emphasize the importance of the original features and stabilize the model training, we also fuse the original features with the aggregated features through skip connections, as shown in Eq. 5 and 10. Finally, to mine the cross-modal interactions, we construct a cross-modal heterogeneous graph with image patches and claim tokens as nodes, and design a GAA layer to aggregate features. The uni-modal features and multi-modal features are concatenated, which can provide richer clues for the rumor classification task. 2) We take advantage of the embedded text information. Using OCR technology, we extract the text content first, and then design the SA and GA modules to adaptively filter which word is more important. 3) Our proposed MGIN-AG framework is composed of multiple independent modules, and adopts the concatenation approach for modality fusion. However, we can see from Figure 2 that the dominant modality may inhibit the optimization of the weak modality. Therefore, we propose the MOE strategy on the basis of OGM to alleviate the possible under-optimization problems of each module in the framework.

Table 2: The performance of MGIN-AG and baselines on Chinese Weibo where R and N refer to Rumor and Non-rumor.

Weibo					
Method	Class	Acc.	Prec.	Rec.	F_1
ViT	R	0.684	0.669	0.665	0.641
	N		0.673	0.684	0.658
BERT	R	0.911	0.919	0.883	0.886
	N		0.890	**0.932**	0.899
EANN	R	0.818	0.785	0.879	0.821
	N		0.860	0.759	0.795
MVAE	R	0.851	0.869	0.835	0.840
	N		0.842	0.869	0.845
HMCAN	R	0.922	0.914	0.914	0.905
	N		0.904	0.897	0.891
MGIN-AG	R	**0.940**	**0.934**	**0.941**	**0.929**
	N		**0.928**	0.931	**0.922**

Table 3: The performance of MGIN-AG and baselines on English Pheme where R and N refer to Rumor and Non-rumor.

PHEME					
Method	Class	Acc.	Prec.	Rec.	F_1
ViT	R	0.771	0.760	0.791	0.769
	N		0.783	0.757	0.761
BERT	R	0.863	**0.885**	0.857	0.867
	N		0.849	0.862	0.851
EANN	R	0.776	0.757	0.810	0.776
	N		0.796	0.758	0.768
MVAE	R	0.833	0.802	0.888	0.839
	N		0.888	0.775	0.820
HMCAN	R	0.866	0.880	0.849	0.862
	N		0.850	**0.870**	0.858
MGIN-AG	R	**0.876**	0.868	**0.889**	**0.874**
	N		**0.892**	0.858	**0.870**

4.5 Ablation Study

To verify the effectiveness of the different modules of MGIN-AG, we compare it with the following variants:

- **w/o V**: MGIN-AG without the visual information.
- **w/o C**: MGIN-AG without the claim, which only relies on images and embedded text to classify rumors.
- **w/o T**: MGIN-AG without the embedded text information from images.

Table 4: Results of ablation study on the PHEME and Weibo.

Model	PHEME		Weibo	
	Acc.	F_1	Acc.	F_1
MGIN-AG	**0.876**	**0.872**	**0.940**	**0.926**
w/o V	0.865	0.860	0.936	0.923
w/o C	0.776	0.768	0.844	0.815
w/o T	0.867	0.862	0.920	0.907
w/o CM	0.869	0.867	0.933	0.920
w/o MOE	0.870	0.866	0.927	0.909

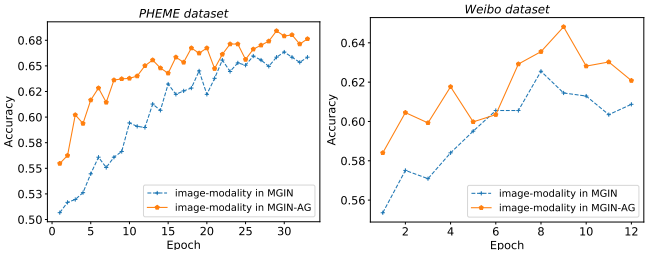

Figure 4: Performance of image modality in MGIN and MGIN-AG, where MGIN removes MOE.

- **w/o CM**: We remove the interaction module between the image and the claim.
- **w/o MOE**: We remove the MOE training strategy, that is, the MGIN-AG model cannot adaptively adjust the optimization process of each module.

The experimental results are shown in Table 4 and we can observe that:

1) Visual information, claim information and embedded text are all important clues, and their corresponding ablation variants perform worse than the complete MGIN-AG. It should be noted that since multi-modal rumor detection is dominated by claims, the performance of the model will decline significantly when the claim information is not considered.
2) The lack of the interaction module between the image and the claim will reduce the overall performance of MGIN-AG. Since the model does not learn the relation between image patches and claim tokens, and in particular cannot implicitly mine their mutual enhancement and consistency.
3) The introduction of the MOE strategy helps to improve the model performance, mainly because it can effectively balance the optimization speed of different modules in the MGIN-AG framework. In addition, since the image quality in the Weibo dataset is better than that in PHEME, that is, Weibo images cover more information, the MOE strategy is easier to help the model obtain higher test accuracy in Weibo.

4.6 Modality Optimization Analysis

To further verify whether MOE can help to improve the under-optimization phenomenon of weak modality in MGIN-AG framework, we demonstrate the performance of image modality in MGIN and MGIN-AG based on the PHEME and Weibo datasets, respectively. It can be observed in Figure 4 that when the MOE training strategy is adopted, the performance of the image modality is improved in all the training epochs, proving that the introduction of MOE is beneficial.

(a) Breaking news! It's illegal to sell dog meat!

(b) They are very poor. The school children really need clothes because they are too poor to afford them.

(c) A drunk man driving a black Porsche crashed into a parked school bus.

Figure 5: Illustration of some typical cases detected by MGIN-AG.

4.7 Qualitative Evaluation

To illustrate the effectiveness of our MGIN-AG, we give three representative cases, all of which have been successfully classified by the model. It can be seen that, in Figure 5(a), the image modality is simple and normal, but the description of the claim is exaggerated and suspicious, which provides a useful clue to the model. In Figure 5(b), the claim looks normal, but the visually striking image doesn't match the text. Using attention-grabbing fake images as illustrations for claims is a common ploy among rumor-makers, but our model also successfully identifies it. In Figure 5(c), our model not only encodes some important areas of the image and key entity clue words such as "black Porsche" and "bus" in the claim, but also measures the interactivity and consistency between them for multi-modal rumor detection.

5 CONCLUSION

In this paper, we propose a novel multi-modal rumor detection framework MGIN-AG. First, to learn the correlation between image patches or claim tokens, we construct two homogeneous in-modal graphs based on cosine similarity and dependency tree, respectively. Then, we consider the interaction between the modalities and a heterogeneous cross-modal graph is constructed in which signed attention mechanism is used to capture mutual enhancement or consistency between image patches and claim tokens. Next, the embedded text in the image is also regarded as one of the important clues to classify rumors. Finally, the MOE strategy is introduced to balance the optimization process of modules. Experimental results on English Pheme and Chinese Weibo show that our MGIN-AG outperforms the SOTA baselines.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Nos.62276177, 62006167 and 61836007), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 549–556.

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

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 81–89.

[6] Z. He, C. Li, F. Zhou, and Y. Yang. 2021. Rumor Detection on Social Media with Event Augmentations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[7] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[8] Nikhil Ketkar. 2017. Introduction to pytorch. In *Deep learning with python*. Springer, 195–208.

[9] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[10] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[11] Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2019. Fake news: Evidence from financial markets. *Available at SSRN 3237763* (2019).

[12] Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1173–1179.

[13] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648* (2020).

[14] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).

[15] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*. 585–593.

[16] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*. 3049–3055.

[17] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8238–8247.

[18] Lahari Poddar, Wynne Hsu, Mong Li Lee, and Shruti Subramaniyam. 2018. Predicting stances in twitter conversations for detecting veracity of rumors: A neural approach. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 65–72.

[19] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[20] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[21] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709* 8 (2017).

[22] Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1412–1423.

[23] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. In *Proceedings of the ACM Web Conference 2022*. 2789–2797.

[24] Tian Tian, Yudong Liu, Xiaoyu Yang, Yuefei Lyu, Xi Zhang, and Binxing Fang. 2020. QSAN: A quantum-probability based signed attention network for explainable false information detection. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1445–1454.

[25] NGUYEN VAN HA, K Sugiyama, P Nakov, and MY Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. (2020).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[27] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12695–12705.

[28] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[29] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 540–547.

[30] Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. *arXiv preprint arXiv:1909.08211* (2019).

[31] Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision tree-based co-attention networks for explainable claim verification. *arXiv preprint arXiv:2004.13455* (2020).

[32] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.

[33] Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A State-independent and Time-evolving Network for Early Rumor Detection in Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. 9042–9051.

[34] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A Convolutional Approach for Misinformation Identification.. In *IJCAI*. 3901–3907.

[35] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 796–805.

[36] Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multi-modal meta multi-task learning for social media rumor detection. *IEEE Transactions on Multimedia* 24 (2021), 1449–1459.

[37] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. In *IJCAI*. 2413–2419.