MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection

Jiaqi Zheng[1] , **Xi Zhang**[1*] , **Sanchuan Guo**[1] , **Quan Wang**[1] ,
Wenyu Zang[2] and **Yongdong Zhang**[3,1]
[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
[2]China Electronics Corporation
[3]University of Science and Technology of China
{zjq1, zhangx, guosc, wangquan}@bupt.edu.cn, wenyuzang@sina.com, zhyd73@ustc.edu.cn

Abstract

Rumor spreaders are increasingly taking advantage of multimedia content to attract and mislead news consumers on social media. Although recent multimedia rumor detection models have exploited both textual and visual features for classification, they do not integrate the social structure features simultaneously, which have shown promising performance for rumor identification. It is challenging to combine the heterogeneous multi-modal data in consideration of their complex relationships. In this work, we propose a novel Multi-modal Feature-enhanced Attention Networks (MFAN) for rumor detection, which makes the first attempt to integrate textual, visual, and social graph features in one unified framework. Specifically, it considers both the complement and alignment relationships between different modalities to achieve better fusion. Moreover, it takes into account the incomplete links in the social network data due to data collection constraints and proposes to infer hidden links to learn better social graph features. The experimental results show that MFAN can detect rumors effectively and outperform state-of-the-art methods.

1 Introduction

With the rapid development of social media such as Twitter and Weibo, rumors can quickly spread over these platforms, which can lead to significant negative impacts on society. For example, the rumor blaming 5G for the coronavirus pandemic had led to arson attacks on more than 70 cell phone towers in the UK in 2020[1]. Due to the large amounts of user-generated content every day, it is desirable to automatically identify rumors to minimize the harmful impacts.

Traditional rumor detection models mainly rely on extracting textual features as source post representations for classification, either with traditional learning models such as decision trees [Castillo *et al.*, 2011] or deep neural networks (DNN) based models such as recurrent neural networks (RNN) and convolutional neural networks (CNN) [Ma *et al.*, 2016; Yu *et al.*, 2017]. With the prevalent of multimedia posts on social media, rumor spreaders tend to utilize visual content together with textual content to attract more attention and get rapid dissemination. To address this issue, a line of multimedia rumor detectors have been proposed to fuse textual and visual features based on DNN to produce multi-modal post representations, which have shown better performance than solely using the textual data [Khattar *et al.*, 2019; Wang *et al.*, 2018; Zhou *et al.*, 2020]. However, one common limitation of these studies is that they didn't consider the graphical social contexts simultaneously, which have been proved to be beneficial to improve the detection performance [Yuan *et al.*, 2019].

The social context of a source post commonly involves its forwarding users and the corresponding comments. Based on these entities and their connections, a heterogeneous graph can be constructed to model the structure information. Then graphical models such as graph attention networks (GAT) [Veličković *et al.*, 2017] and graph convolutional networks (GCN) [Kipf and Welling, 2016] can be utilized to aggregate adjacent node information to obtain better node representations for rumor detection [Yuan *et al.*, 2019; Yang *et al.*, 2021]. With the help of graphical models, connected instances can exchange information and facilitate each other's learning. However, the existing graph-based detectors suffer from several limitations: (1) the quality of node representation learning depends highly on reliable links among entities. Due to the privacy issue or data crawling constraint, the available social graph data is very likely to lack some important links among entities. Therefore, it is necessary to complement latent links on the social graph to achieve a more accurate detection; (2) there may be various latent relations between adjacent nodes on a graph, while the conventional neighborhood aggregation procedure of graph neural networks (GNN) may not be able to differentiate their effects on the representation of a target node, leading to inferior performance; (3) how to effectively integrate the learned social graph features with other modality features (e.g., visual features) is less explored in existing studies.

To address the above challenges, we propose a novel Multi-modal Feature-enhanced Attention Network (MFAN) for multimedia rumor detection. We make the first attempt to jointly model textual, visual, and social graph features in one

*Corresponding author

[1]https://www.cnet.com/health/5g-coronavirus-conspiracy-theory-sees-77-mobile-towers-burned-report-says/

framework. Given the multimedia post features and the social graph features, a straightforward solution is to adopt the attention mechanism between the two groups of features and effectively aggregate them to produce distinctive features for rumor detection. In this work, we improve the multi-modal fusing mechanism by considering the cross-modal semantic alignment. Specifically, a self-supervised loss is introduced to align the source post representations learned from two distinct views, i.e., the textual-visual view and the social graph view, aiming to improve the representation learning in each view. In addition, to obtain better social graph structures, we improve the graph representation learning from two perspectives. On the one hand, we propose to infer potential links between nodes in the social graph to alleviate the incomplete link issue. On the other hand, we utilize a signed attention mechanism to capture both positive and negative neighborhood correlations to achieve better node representations. Through the above enhanced cross-modal fusion and social graph representation learning, we can promote the performance of multimedia rumor detection.

The main contributions can be summarized as follows:

- We propose a multi-modal feature-enhanced attention network for multimedia rumor detection, which can effectively combine textual, visual, and social graph features in one unified framework.
- We introduce a self-supervised loss to align the source post representations in different views to achieve better multi-modal fusion.
- We improve the social graph feature learning by enhancing both the graph topology and neighborhood aggregation procedure.
- We empirically show that the proposed model can effectively identify rumors and outperform the state-of-the-art baselines on two large-scale real-world datasets.

2 Related Work

Early rumor detection methods usually manually extract features from text content for classification [Castillo *et al.*, 2011; Popat, 2017], which requires massive human efforts. More recently, DNN-based methods [Ma *et al.*, 2016; Karimi *et al.*, 2018; Ma *et al.*, 2018] have been proposed to learn the post representations to promote the detection performance. However, these studies don't consider the visual features and social context features that would be beneficial.

Multi-modal data have been exploited by a set of studies to facilitate the detection. [Khattar *et al.*, 2019] uses bidirectional long short-term memory (Bi-LSTM) to extract textual and visual representations. [Wang *et al.*, 2018] jointly considers textual and visual information and removes the event-specific features. [Yuan *et al.*, 2019] combines text and heterogeneous graph structures for rumor detection. However, they adopt simple fusion methods and may not be able to capture the complex relationships among multi-modal data.

To explore the rich correlations between different modalities, [Qian *et al.*, 2021] jointly models the multi-modal context information and the hierarchical semantics of text in a unified model. [Lu and Li, 2020] learns the representations of user interactions, retweet propagation, and their correlation with source tweets. [Wu *et al.*, 2021] stacks multiple co-attention layers to fuse the multi-modal features. [Jin *et al.*, 2017] jointly considers the fusion of text, image, and social context information but ignores the social graph structure.

Graph-based rumor detectors have been proposed to exploit the social graph to enhance the post representations [Yuan *et al.*, 2019; Yang *et al.*, 2021]. However, they ignore the incomplete link issue in the dataset. Although [Wei *et al.*, 2021] has considered the propagation uncertainty between different nodes, it doesn't complement the missing links on the social graph nor conduct the multi-modal fusion to promote the representation learning.

A comparison between our work and the related studies is shown in Table 1. The uniqueness of our work lies in: jointly using textual, visual, and social graph features, involving multi-modal alignment for better fusion, and utilizing potential relationships to enhance the graph features.

	TF	VF	SF	PR	PN	MI	MA	AL
[Wang *et al.*, 2018]	✓	✓						
[Khattar *et al.*, 2019]	✓	✓						
[Tian *et al.*, 2020]	✓				✓			
[Zhou *et al.*, 2020]	✓	✓				✓		
[Wei *et al.*, 2021]	✓		✓	✓				
[Yuan *et al.*, 2019]	✓		✓					
Our Work	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: A comparison of related studies. Column notations: Textual Feature (TF), Visual Feature (VF), Social Graph Feature (SF), Potential Relationship in Social Networks (PR), Positive and Negative Relationship (PN), Multi-modal Interaction (MI), Multi-modal Alignment (MA), Adversarial Learning (AL).

3 Problem Definition

Let $P = \{p_1, p_2, \cdots, p_n\}$ be a set of multimedia posts on social media with both texts and images. For each post $p_i \in P$, $p_i = \{t_i, v_i, u_i, c_i\}$, where t_i, v_i and u_i denote the text, image and user who have published the post. $c_i = \{c_i^1, c_i^2, \cdots, c_i^j\}$ represents the set of comments of p_i. Moreover, each comment c_i^j is posted by a corresponding user u_i^j.

In order to represent user behaviors on social media, we establish a graph $G = \{V, A, E\}$, where V is a set of nodes, including user nodes, comment nodes, and post nodes. $A \in \{0,1\}^{|V|*|V|}$ is an adjacency matrix between nodes to describe the relationships between nodes, including posting, commenting, and forwarding. E is the set of edges.

We define rumor detection as a binary classification task. $y \in \{0, 1\}$ denotes class labels, where $y = 1$ indicates rumor, and $y = 0$ otherwise. Our goal is to learn the function $F(p_i) = y$ to predict the label of a given post p_i.

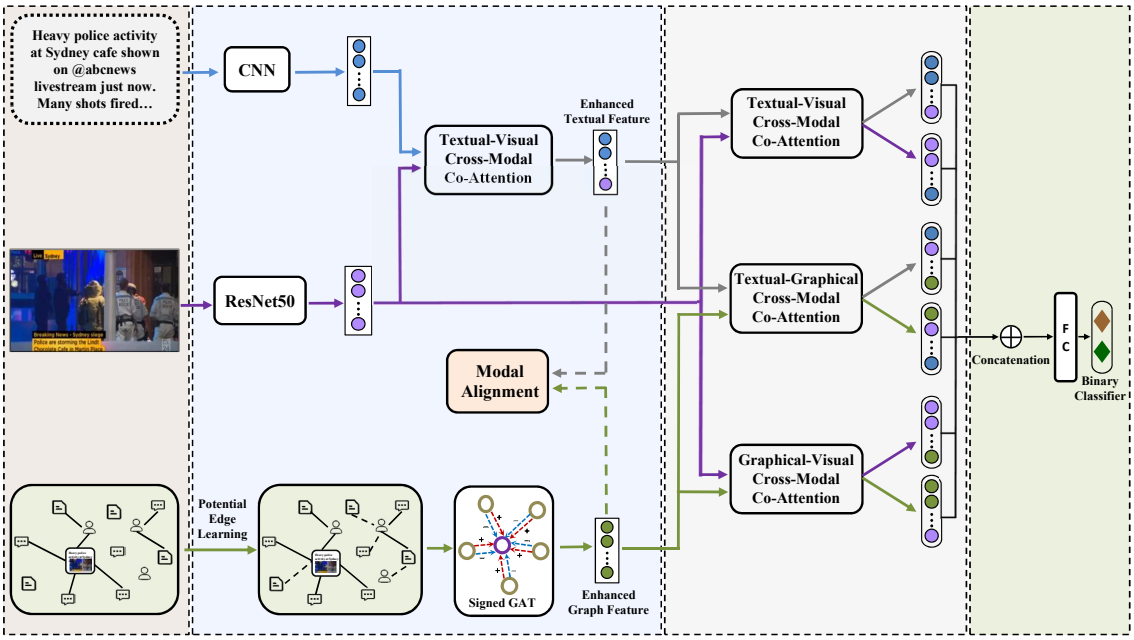

Figure 1: The proposed framework MFAN. We first obtain the three modal features of textual, visual, and graphical for a post on social media through feature extractors. Then we use the visual feature to enhance the textual feature and utilize the potential relationship in social networks to enhance the graphical feature. We perform the modal alignment between the above two enhanced features. The cross-modal co-attention mechanism is used to obtain the enhanced features between every two modalities. Then we integrate all the enhanced modal features for rumor detection.

4 Methodology

The focus of our proposal is to effectively combine textual, visual, and social graph features to improve the rumor detection. To this end, we first extract the three types of features. In order to produce better social graph features, we propose enhancing both the graph topology and aggregation procedure based on GAT. We then capture cross-modal interaction and alignment to achieve better multi-modal fusion. Finally, we concatenate the enhanced multi-modal features for classification. We also apply adversarial training to promote the robustness. The overall architecture is illustrated in Figure 1.

4.1 Textual and Visual Feature Extractor

Textual Representations

We employ CNN with pooling to extract the semantic feature of sentences. Firstly, for each post p_i, its text t_i is padded or truncated to have the same number of tokens, i.e., L, which is represented as

$$\mathcal{O}^i_{1:L} = \{o^i_1, o^i_2, \cdots, o^i_L\} \tag{1}$$

where $o \in \mathbb{R}^d$, d is the dimension of word embeddings and o^i_j denotes the word embedding of the j-th word of t_i.

Then, we apply convolution layer on the word embedding matrix $\mathcal{O}^i_{j:j+k-1}$ to get the feature map s_{ij}, where k is the size of the receptive field. We denote $s^i = \{s_{i1}, s_{i2}, \cdots, s_{i(L-k+1)}\}$. Then, we use max pooling over s^i to obtain $\hat{s^i} = max(s^i)$. We use $d/3$ filters with varying receptive filed $k \in \{3, 4, 5\}$ to obtain semantic features of different granularities.

Finally, we concatenate all filters' outputs to form the overall textual feature of t_i:

$$R^i_t = concat(\hat{s^i_{k=3}}, \hat{s^i_{k=4}}, \hat{s^i_{k=5}}) \tag{2}$$

Visual Representations

We use the pre-trained model ResNet50 [He *et al.*, 2016] trained over the ImageNet database to extract image v_i's feature. Firstly, we extract the output of the second last layer of ResNet50 and denote it as V^i_r. Then, we pass it through a fully connected layer to obtain the final visual feature with the same dimension as the textual feature, that is,

$$R^i_v = \sigma(W_v * V^i_r). \tag{3}$$

where W_v is the weight matrix of the fully connected layer and $\sigma(\cdot)$ is an active function such as $sigmoid$.

4.2 Enhanced Social Graph Feature Learning

Inferring Hidden Links

To alleviate the missing link issue, we propose to infer the hidden links between nodes in social networks. According to network homophily, similar nodes may be more likely to attach to each other than dissimilar ones. We thus calculate the feature similarity between different nodes and infer links between nodes with high similarity. Specifically, we define the node embedding matrix as $X \in \mathbb{R}^{|V| \times d}$, where d is the dimension size. There are three types of nodes in X, we use the sentence vectors as the initial embeddings of the post and comment nodes, and use the average value of the post node embeddings posted by the user as the initial user embedding.

Then we calculate the correlation β_{ij} between node n_i and n_j based on their cosine similarity as

$$\beta_{ij} = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|} \tag{4}$$

where x_i and x_j are node embeddings of n_i and n_j. We then infer there exists a potential edge between them if the similarity is above 0.5, that is,

$$e_{ij} = \begin{cases} 0\ ,\ if\ \beta_{ij}<0.5 \\ 1\ ,\ otherwise \end{cases} \tag{5}$$

Then we enhance the original adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ with the inferred potential edges. a_{ij} denotes the element of A, where $a_{ij} = 1$ indicates there is an edge between n_i and n_j and $a_{ij} = 0$ otherwise. Then the element a'_{ij} of the enhanced adjacency matrix $A^{'}$ is defined as

$$a_{ij}^{'} = \begin{cases} 0\ ,\ if\ e_{ij} = 0\ and\ a_{ij} = 0 \\ 1\ ,\ otherwise \end{cases} \tag{6}$$

Capturing Multi-aspect Neighborhood Relations

Here we aim to capture the social graph structure information with GAT. Different from conventional GAT, we introduce the signed attention mechanism to capture the positive and negative correlations between neighboring nodes to obtain better graph features. We start from the vanilla GAT [Veličković *et al.*, 2017] and then propose our signed attention based GAT to capture the multi-aspect correlations.

The key of GAT is the aggregation of the neighborhood information. For node n_i and its neighbor node set $\mathcal{N}_i = \left\{n_1^{'}, n_2^{'}, \cdots, n_{|\mathcal{N}_i|}^{'}\right\}$, we first calculate the set of attention weights $\mathcal{E}_i = \left\{e_{i1}^{'}, e_{i2}^{'}, \cdots, e_{i|\mathcal{N}_i|}^{'}\right\}$ between n_i and each node in $\mathcal{N}_i$ by

$$e_{ij}^{'} = LeakyReLU(\hat{a}\left[Wx_i \| Wx_j^{'}\right]) \tag{7}$$

where $\|$ means concatenation operation, $\hat{a}$ and W are learnable parameters, x_i and $x_j^{'}$ denote the node embeddings of node n_i and $n_j^{'}$, $j \in \{1, 2, \cdots, |\mathcal{N}_i|\}$ and $n_j^{'} \in \mathcal{N}_i$.

Then, we use the softmax function to perform the weight normalization operation on the attention weights. The attention weight $e_{ij}^{'}$ between n_i and $n_j^{'}$ may be a negative value, which will become extremely small after softmax function. In fact, the attention weight between nodes contains a potential positive and negative relationship, which will be ignored by directly using the softmax function. For example, for a specific node n_t, we obtain its weights with the neighbor nodes as $\mathcal{E}_t = \{0.7, 0.3, -0.1, -0.9\}$. After normalization by the softmax function, the weights become $\mathcal{E}_t^{'} = \{0.43, 0.29, 0.20, 0.09\}$. It can be seen that the node corresponding to the value "-0.9" in the weight vector has the smallest contribution to the output. However, "-0.9" may indicate that the two node vectors are in opposite directions. This kind of large negative relations may also be beneficial for rumor detection. For instance, it may reflect camouflage behaviors such as a rumor spreader buying some honest users as fans or a comment opposing a source post [Yang *et al.*, 2021], and their node vectors can be intrinsically negatively correlated. Unfortunately, existing GATs ignore such negative correlations.

To address this issue, inspired by QSAN [Tian *et al.*, 2020], we design a signed attention based GAT, namely Signed GAT, which uses signed attention to involve both the positive and negative relationships between nodes. Specifically, for node n_i, we denote the inversion of the attention weights $\mathcal{E}_i$ of its neighbor nodes as $\tilde{\mathcal{E}}_i = -\mathcal{E}_i$. We then calculate the normalized weights for both $\mathcal{E}_i$ and $\tilde{\mathcal{E}}_i$ with the softmax function,

$$\begin{aligned} \mathcal{E}_i^{'} &= softmax(\mathcal{E}_i) \\ \tilde{\mathcal{E}}_i^{'} &= softmax(\tilde{\mathcal{E}}_i) \end{aligned} \tag{8}$$

In order to capture both positive and negative relations between nodes, we utilize $\mathcal{E}_i^{'}$ and $-\tilde{\mathcal{E}}_i^{'}$ respectively to obtain the weighted sum of the neighbor nodes' features. Then we concatenate the two vectors together and pass it through a full connected layer to obtain the final node feature. For instance, the node feature of n_i can be obtained by

$$\hat{x_i} = \sigma(W_n * (\mathcal{E}_i^{'} * X_j \| - \tilde{\mathcal{E}}_i^{'} * X_j)) \tag{9}$$

where W_n is the weight matrix of the fully connected layer, $\sigma(\cdot)$ is an active function and X_j is the feature matrix of $\mathcal{N}_i$.

Graph Feature Extractor

We then introduce how to obtain the social graph feature based on the enhanced social graph and the signed GAT. Firstly, we enhance the original social graph by augmenting the inferred potential edges, and initialize three types of nodes in the graph. For post and comment nodes, we use their textual features as the initial embeddings. For user nodes, we use the average of their post and comment embeddings as the initial embeddings to reflect the user characteristics.

Then we use Signed GAT to extract graph structure features from the enhanced social graph. For each node, we update its embedding according to Eqn. (9) and obtain the updated node embedding matrix $\hat{X} \in \mathbb{R}^{|V|\times d}$, where $|V|$ is the number of nodes and d is the dimension size. Then a multi-head attention mechanism [Vaswani *et al.*, 2017] is adopted to capture features from different perspectives. We concatenate the updated node embeddings of each head together as the overall graph feature:

$$\hat{G} = \mathop{\Big\|}_{h=1}^{H} \sigma(\hat{X}_h) \tag{10}$$

where H denotes the number of heads. Then the graph feature R_g^i of the i-th post p_i corresponds to the i-th column of $\hat{G}$.

4.3 Multi-modal Feature Fusing

In this work, as there are three types of modalities, we adopt a hierarchical fusion schema with the co-attention method [Lu *et al.*, 2019]. In order to capture different aspects of cross-modal relationships and enhance the multi-modal features, we propose to enforce the cross-modal alignment with a self-supervised loss.

Cross-modal Co-attention Mechanism

We use the co-attention mechanism to capture the mutual information between different modalities. It learns the attention weights between different modal features to enhance the cross-modal feature.

Specifically, for each modal, we first use multi-head self-attention [Vaswani *et al.*, 2017] to enhance the intra-modal feature representation. For example, for the text feature R_t^i, we use $Q_t^i = R_t^i W_t^Q$, $K_t^i = R_t^i W_t^K$, and $V_t^i = R_t^i W_t^V$ to calculate its query matrix, key matrix and value matrix respectively, where $W_t^Q, W_t^K, W_t^V \in \mathbb{R}^{d \times \frac{d}{H}}$ are linear transformations and H is the number of heads. We then produce the multi-head self-attention feature of the text modal as

$$Z_t^i = (\mathop{\Vert}_{h=1}^{H} softmax(\frac{Q_t^i {K_t^i}^T}{\sqrt{d}})V_t^i)W_t^O \tag{11}$$

where h denotes the h-th head, and $W_t^O \in \mathbb{R}^{d \times d}$ is the output linear transformations. We perform the same operations on R_v^i and R_g^i to obtain the corresponding features Z_v^i and Z_g^i.

Then we use the co-attention mechanism to produce the enhanced multi-modal features. Specifically, in order to perform the textual-visual co-attention for p_i, we first perform a similar operation as the above self-attention, but replace R_t^i with Z_v^i to get the query matrix Q_v^i, and replace R_t^i with Z_t^i to get the key matrix K_t^i and value matrix V_t^i. Then we obtain the cross-modal enhanced feature Z_{vt}^i as

$$Z_{vt}^i = (\mathop{\Vert}_{h=1}^{H} softmax(\frac{Q_v^i {K_t^i}^T}{\sqrt{d}})V_t^i)W_{vt}^O \tag{12}$$

where $W_{vt}^O \in \mathbb{R}^{d \times d}$ is the output linear transformations.

Note that Z_{vt}^i represents the enhanced textual feature with the visual feature based on their correlations. Based on the same co-attention procedure, we can obtain the enhanced visual feature Z_{tv}^i with the textual feature by exchanging the roles of the two modalities in Eqn. (12).

Multi-modal Alignment

Based on the co-attention mechanism, we can obtain the enhanced textual feature with the visual feature and vice versa. However, for the source post, its representations in different modalities should be intrinsically related. Such inter-modal correspondences are not covered by the co-attention mechanism. We thus introduce the modal alignment via enforcing the enhanced textual feature of the post close to its enhanced graphical features in order to refine the representations learned in each modality.

Specifically, for a post p_i, its enhanced graph feature Z_g^i and enhanced textual feature Z_{vt}^i are transformed into the same modal feature space, that is,

$$\begin{aligned} Z_g^{i\prime} &= W_g{}' Z_g^i \\ Z_t^{i\prime} &= W_t{}' Z_{vt}^i \end{aligned} \tag{13}$$

where $W_g{}'$ and $W_t{}'$ are learnable parameters. Then we narrow the distance between $Z_g^{i\prime}$ and $Z_t^{i\prime}$ with the MSE loss for modal alignment:

$$\mathcal{L}_{align} = \frac{1}{n}\sum_{i=1}^{n}(Z_g^{i\prime} - Z_t^{i\prime})^2 \tag{14}$$

Statistic	Non-rumors	False Rumors	Images	Users	Comments
PHEME	1428	590	2018	894	7388
Weibo	877	590	1467	985	4534

Table 2: The statistics of two datasets.

where n is the total number of posts. We then get the alignment-refined textual feature $\tilde{Z}_t^i$ and graphical feature $\tilde{Z}_g^i$, which is used for the following multi-modal fusion.

Fusing the Above Multi-modal Features

We again perform the aforementioned cross-modal co-attention mechanism among each pair of the three modal features, i.e., $\tilde{Z}_t^i$, $\tilde{Z}_g^i$ and Z_v^i, and finally get six cross-modal enhanced features: $\tilde{Z}_{tv}^i$, $\tilde{Z}_{vt}^i$, $\tilde{Z}_{gt}^i$, $\tilde{Z}_{tg}^i$, $\tilde{Z}_{gv}^i$, and $\tilde{Z}_{vg}^i$. We then concatenate them as the final multi-modal feature:

$$Z^i = concat(\tilde{Z}_{tv}^i, \tilde{Z}_{vt}^i, \tilde{Z}_{gt}^i, \tilde{Z}_{tg}^i, \tilde{Z}_{gv}^i, \tilde{Z}_{vg}^i) \tag{15}$$

4.4 Classification with Adversarial Training

We feed the final multi-modal feature Z^i of post p_i into the fully connected layer to predict whether p_i is a rumor or not,

$$\hat{y}_i = softmax(W_c Z^i + b) \tag{16}$$

where $\hat{y}_i$ denotes the predicted probability of p_i being a rumor. Then we use the cross-entropy loss function as

$$\mathcal{L}_{classify} = -y\log(\hat{y}_i) - (1-y)\log(1-\hat{y}_i) \tag{17}$$

The final loss can be written as follows:

$$\mathcal{L} = \lambda_c \mathcal{L}_{classify} + \lambda_a \mathcal{L}_{align} \tag{18}$$

where λ_c and λ_a are used to balance the two losses.

As the text contents in social media may not follow the strict grammar rules, in order to adapt to such grammatical irregularity, we add adversarial perturbations at the text embedding level to enhance the robustness of the model. We use PGD [Madry *et al.*, 2017], a widely used adversarial training method. Specifically, we calculate the gradient for the textual feature in each training iteration and use it to calculate the adversarial perturbation, which is added to the textual feature. We then recalculate the gradient on the updated textual feature. We repeat this process k times and use a spherical space to limit the extent of the perturbation. Finally, the above adversarial gradients are accumulated to the original gradient, which is then used for parameter updating. More details can refer to [Madry *et al.*, 2017].

5 Experiments

5.1 Datasets

We evaluate our model on two real-world datasets: Weibo [Song *et al.*, 2019] and PHEME [Zubiaga *et al.*, 2017]. The Weibo dataset is collected from the most popular social media in China. PHEME is constituted by tweets on the Twitter platform and based on five breaking news. Each dataset contains texts, the attached images, and comments. In this work, we focus on detecting rumors with three modal features, i.e., the textual, visual, and social graph features. Thus, we remove the data instances without any text or image. Table 2 shows the statistics of the resulting two datasets after removal.

	PHEME				Weibo			
Method	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
EANN	77.13±0.96	71.39±1.07	70.07±2.19	70.44±1.69	80.96±2.26	80.19±2.37	79.68±2.46	79.87±2.40
MVAE	77.62±0.64	73.49±0.81	72.25±0.90	72.77±0.81	71.67±0.89	70.52±0.95	70.21±1.01	70.34±0.98
QSAN	75.13±1.19	69.97±2.03	65.80±1.72	66.87±1.70	71.01±1.81	71.02±0.95	67.54±3.27	67.58±3.59
SAFE	81.49±0.84	79.88±1.22	79.50±0.81	79.68±0.70	84.95±0.85	84.98±0.82	84.95±0.91	84.96±0.86
EBGCN	82.99±0.65	81.31±0.73	79.29±0.71	79.82±0.64	83.14±2.01	85.46±2.12	81.76±1.54	81.45±1.74
GLAN	83.32±1.64	81.25±2.06	77.13±3.26	78.51±2.68	82.44±2.02	82.45±2.26	80.86±1.71	81.26±1.93
MFAN	**88.73±0.83**	**87.07±1.41**	**85.61±1.65**	**86.16±1.04**	**88.95±1.43**	**88.91±1.60**	**88.13±1.68**	**88.33±1.53**

Table 3: Results of comparison among different models on PHEME and Weibo datasets.

5.2 Baselines

We compare our model with the following strong baselines:

- **EANN** [Wang *et al.*, 2018] is a GAN-based model exploiting both text and image data. It derives event-invariant features and benefits newly arrived events.
- **MVAE** [Khattar *et al.*, 2019] uses a bimodal variational autoencoder coupled with a binary classifier for multi-modal fake news detection.
- **QSAN** [Tian *et al.*, 2020] integrates the quantum-driven text encoding and a novel signed attention mechanism for false information detection.
- **SAFE** [Zhou *et al.*, 2020] jointly exploits multi-modal features and cross-modal similarity to learn the representation of news articles.
- **EBGCN** [Wei *et al.*, 2021] rethinks the reliability of latent relations in the propagation structure by adopting a Bayesian approach.
- **GLAN** [Yuan *et al.*, 2019] jointly encodes the local semantic and global structural information and applies a global-local attention network for rumor detection.

Among them, EANN, MVAE and SAFE exploit both textual and visual data. QSAN only exploits the textual data. Social graphical features are considered by EBGCN and GLAN. None of them consider data from all three modalities like our proposed model.

5.3 Implementation Details

We split the datasets for training, validation, and testing with a ratio of 7:1:2. The evaluation metrics include Accuracy, Precision, Recall, and F1. We use word vectors provided in [Yuan *et al.*, 2019] as initialized word embeddings. The number of heads H is set to 8. λ_c and λ_a are set to 2.15 and 1.55. We choose the best parameter configuration based on the performance of the proposed model. We use Adam [Kingma and Ba, 2014] to optimize our objective function. The learning rate used in the training process is 0.002. We perform 5 runs throughout all experiments and report the average results and standard deviation results.

5.4 Results and Discussion

Table 3 shows the performance of the comparison methods. On both datasets, our model MFAN significantly outperforms all the other approaches in all the metrics. GLAN and EBGCN outperform most other methods, indicating that the social graph information is beneficial for rumor detection. For methods that consider both textual and visual information, SAFE outperforms other methods, indicating the importance of considering interactions between modalities. Our MFAN outperforms GLAN and EBGCN, demonstrating that considering visual data, latent links, and modal alignment can further improve detection performance.

Method		-w/o V	-w/o G	-w/o P	-w/o A	MFAN
PHEME	Acc.	85.66	86.29	86.91	87.12	**88.73**
	F1.	82.47	82.15	83.93	84.41	**86.16**
Weibo	Acc.	84.14	85.08	86.17	86.98	**88.95**
	F1.	83.88	84.48	85.44	86.42	**88.33**

Table 4: Experimental results of the variations of MFAN.

5.5 Performance of the Variations

To show the effectiveness of different components in MFAN, we compare it with the sub-models "-w/o V", "-w/o G", "-w/o P", and "-w/o A". They denote the variant of MFAN without considering the visual information, social graph information, potential links, and modal alignment, respectively. The comparison results are shown in Table 4. We can observe that all ablation variants perform worse than the complete MFAN model on both datasets. The results indicate that: (i) visual modal and graph features are both important for rumor detection; (ii) the modal alignment can facilitate the multi-modal fusion; (iii) considering latent links can significantly improve the social graph feature representations.

6 Conclusions

In this paper, we propose a multi-modal rumor detection framework that, for the first time, incorporates three types of modalities, i.e., text, image, and social graph. To improve the social graph feature learning, both the graph topology and neighborhood aggregation procedure are enhanced based on GAT. Our framework enables more effective multi-modal fusion by introducing cross-modal alignment. Evaluations and comparisons on both Chinese and English datasets demonstrate that our model can outperform the state-of-the-art baselines for multimedia rumor detection.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No.61976026, 61902394) and 111 Project (B21049).

References

[Castillo *et al.*, 2011] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *WWW*, pages 675–684, 2011.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM Multimedia*, pages 795–816, 2017.

[Karimi *et al.*, 2018] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *COLING*, pages 1546–1557, 2018.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *WWW*, pages 2915–2921, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

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

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

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

[Ma *et al.*, 2018] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. ACL, 2018.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Popat, 2017] Kashyap Popat. Assessing the credibility of claims on the web. In *WWW Companion*, pages 735–739, 2017.

[Qian *et al.*, 2021] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *SIGIR*, pages 153–162, 2021.

[Song *et al.*, 2019] Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047, 2019.

[Tian *et al.*, 2020] Tian Tian, Yudong Liu, Xiaoyu Yang, Yuefei Lyu, Xi Zhang, and Binxing Fang. Qsan: A quantum-probability based signed attention network for explainable false information detection. In *CIKM*, pages 1445–1454, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *KDD*, pages 849–857, 2018.

[Wei *et al.*, 2021] Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection. *arXiv preprint arXiv:2107.11934*, 2021.

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

[Yang *et al.*, 2021] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. Rumor detection on social media with graph structured adversarial learning. In *IJCAI*, pages 1417–1423, 2021.

[Yu *et al.*, 2017] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. A convolutional approach for misinformation identification. In *IJCAI*, pages 3901–3907, 2017.

[Yuan *et al.*, 2019] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *ICDM*, pages 796–805. IEEE, 2019.

[Zhou *et al.*, 2020] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*, 2020.

[Zubiaga *et al.*, 2017] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*, pages 109–123. Springer, 2017.