A Unified Model for Unimodal and Multimodal Rumor Detection

Haibing Zhou[1], Zhong Qian(✉)[1], Peifeng Li[1], and Qiaoming Zhu[1]

[1]School of Computer Science and Technology, Soochow University, Suzhou, China
hbzhou520@stu.suda.edu.cn, {qianzhong,pfli,qmzhu}@suda.edu.cn

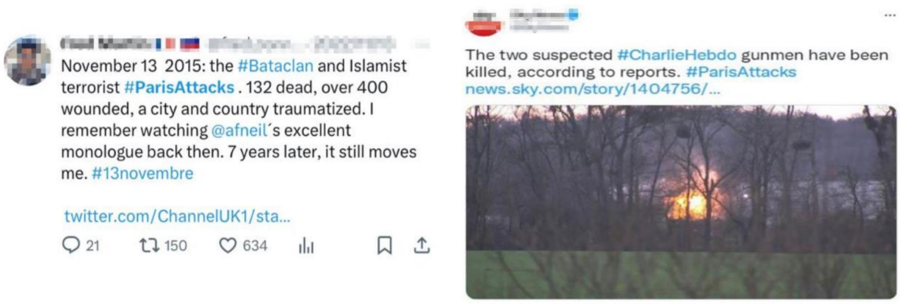
Abstract. Rumor detection aims to determine the truthfulness of a post, no matter it is unimodal (plain text) or multimodal (text and images). However, previous models only considered one of these situations, ignoring the possibility of both occurring simultaneously. Additionally, previous multimodal models often failed to tackle the inconsistency between texts and images, which can produce noise and harm performance. To address the aforementioned issues, we propose a novel unified model for unimodal and multimodal rumor detection, called the Graph Attention Generative Image Network (GAGIN), which is integrated with multimodal alignment. The experimental results on two popular datasets demonstrate that GAGIN outperforms the state-of-the-art baselines.

Keywords: Unified model, Rumor detection, Multimodal rumor detection, Graph attention network, Diffusion model.

1 Introduction

Rumor can lead to serious consequences. For example, during the COVID-19 pandemic, a newly published study shows that approximately 800 people have died due to rumors that drinking high-concentration alcohol can disinfect the body [1]. Rumor detection model can automatically determine whether an event is a rumor and help prevent its dissemination. As shown in **Fig. 1**, rumors can be communicated in plain text or they can be a combination of both visual and textual content. Therefore, previous research can be categorized into unimodal and multimodal methods. Unimodal methods rely on a single type of data, such as text or image, to extract salient features for rumor detection.

Compared to the success of unimodal rumor detection, multimodal approaches are still in its early stages and only focuses on two modalities: text and image. However, there are two issues with multimodal rumor detection. One issue is that previous multimodal approaches are ineffective in detecting unimodal rumors (i.e., those based solely on text), because they heavily rely on both text and image data to extract salient features and their interactions. Using the tweet in **Fig. 1(a)** as an example, it cannot be directly processed by previous multimodal approaches due to the lack of an image, which results in the loss of critical interaction features between text and image.



(a) (b)

Fig. 1. Examples from the social platform Twitter, where (a) is a text-only sample; (b) is a multi-modal sample.

Another issue is the inconsistency between the image and text. Previous approaches often directly concatenate the features of images and texts [2, 3], ignoring the inconsistency between image and text which will harm the performance. For instance, as shown in **Fig. 1(b)**, the text ``The two suspected \#CharlieHebdo gunmen have been killed" does not match the image, which depicts a fire behind some trees.

To address the above two issues, we propose a novel unified model for unimodal and multimodal rumor detection, namely Graph Attention Generative Image Network (GAGIN) which can be applied to text-based and multimodal (i.e., text and image) rumor detection. To address the first issue, we use the advanced diffusion model [4] to generate images based on the text. To solve the second issue, inspired by Clip [5], we compare the similarity between the image generated from the text and the raw image (if it exists) and we encode the visual and text modalities via self-supervised learning for cross-modal alignment to integrate images and texts to detect their inconsistency. Finally, we use texts and images to build graph structures respectively, and use Graph Attention Network (GAT) [7] to obtain the relations between images or texts. The experimental results on two popular datasets show that our GAGIN outperforms the SOTA baselines. The main contributions of this paper are as follows:

1) This paper is the first work to propose an unified model that can be used for both unimodal and multimodal rumor detection, which can benefit from the interaction between text and images that are either original or generated.

2) This paper uses the similarity between the generated image and the raw image to detect the inconsistency of them and utilizes the generated image to resolve the inconsistency.

3) This paper not only considers the difference between the text and the image in the same post, but also learns the relations between texts or images in different posts.

2 Related Work

2.1 Unimodal Rumor Detection

Previous methods usually rely on textual data to extract distinctive features to detect rumors. This type of method uses traditional learning models such as decision trees [7] and support vector machines (SVM) [8] or deep neural network based models. Deep learning models such as RNN [9] and CNN [10] are used to extract high-level text semantics feature representations of text. Due to the popularity of pre-trained models, BERT-based [11] text encoding methods are also adopted [12].

In order to get more useful information from texts, people strive to construct more reasonable neural networks to learn stance-based, emotional, capture comment-based and propagation-based features around metadata, which has achieved satisfactory performance and gained considerable development. Specifically, Wu et al. [13] proposed a sifted multi-task learning model with filtering mechanism to detect fake news by joining stance detection task. Zhang et al. [14] have verified that sentiment signals are differentiated between fake news and real news in their model. Shu et al. utilized both news content and user comments to capture interpretable user comments [15] and proposed a model to study the relations between hierarchical propagation network and rumors for rumor detection [16].

2.2 Multimodal Rumor Detection

These models can not only utilize text information, but also additionally use information other than text (such as images). Specifically, Wang et al. [3] proposed a multimodal model framework where image features encoded by VGG-19 [17] are simply concatenated with text features for rumor detection. Khattar et al. [2] added a decoder based on [3] to improve the quality of multimodal representation. Qian et al. [18] designed a multimodal contextual attention network that can mine hierarchical semantic relationships and model multimodal contextual information for rumor detection. Wu et al. [19] extracted spatial and frequency domain features from images together with text features, and fused them through multiple co-attention modules for rumor detection. On the basis of image and text features, Zheng et al. [20] introduced graph social context features to improve model performance. Sun et al. [21] also introduced graph neural networks and they proposed a fine-grained multimodal graph interaction network that explicitly learns the dependencies between text markers and image patches from a graph perspective and mines the interactions between different modalities for multimedia rumor detection.

The advantages of our study compared with previous work can be summarized as follows. This is the first work to propose a rumor detection model that can be used in either unimodal or multimodal situation simultaneously. Meanwhile, we not only mine the relations between the same modalities, but also learn the relations between different modalities, and can effectively solve the inconsistency between images and texts.

3 Method

3.1 Task Definition

Let $P = \{p_1, p_2, \dots, p_n\}$ be a sequence of posts on social media containing text or both text and images, Since not every post has an image, for each post $p_i \in P$, $p_i = \{t_i, v_i, v_{ti}\}$ or $p_i = \{t_i, v_{ti}\}$, where t_i, v_i and v_{ti} represent the text, image and text to image of p_i. Our goal is to learn a model $f: p_i \rightarrow Y, (p_i \in P)$, to classify each post into the predefined categories $Y = \{0,1\}$, which is the ground-truth label of the post p_i (0/1 denotes non-rumor/rumor).

3.2 Overall Architecture

The architecture of our GAGIN model is shown in **Fig. 2**. We first take out the raw data p_i that needs to be identified of text t_i and image v_i (if it exists) from a post on social media. Secondly, we generate the image v_{ti} from the text, then encode t_i, v_i and v_{ti}, compare the similarity between the encodings of v_{ti} and v_i, and Align the encoding of v_{ti} with the encoding of t_i. Then, we use similarity learning to build graphs and learn the features of the graphs for all texts and images in P. Finally, we use Self-Attention (SA) to further learn all salient features, concatenate them and put them in Fully Connected (FC) layer to distinguish whether P_i is a rumor or not.

3.3 Modules of GAGIN

Raw data feature extraction. The Raw data of the post P_i includes t_i and v_i (if it exists). We first put the text t_i into the pre-trained Diffusion[1] [4] model to generate the image v_{ti}. The process is formulated as follows.

$$v_{ti} = Diffusion(t_i) \tag{1}$$

Then we use pre-trained BERT[2] [11] and Resnet50 [22] to encode the t_i and images v_i (if it exists) or v_{ti}, respectively,

$$R_i^v, R_i^{vt} = Resnet50(\text{images}) \tag{2}$$

$$R_i^t = BERT(t_i) \tag{3}$$

where $R_i^v, R_i^{vt} \in R^d$, $R_i^t \in R^{d'}$, images refers to v_{ti} or v_i.

Multimodal alignment. After obtaining the feature representations R_i^v, R_i^{vt} and R_i^t of v_i, v_{ti} and t_i, we can modally align R_i^{vt} and R_i^t, calculate the similarity between R_i^v and R_i^{vt} to determine whether there is inconsistency between the image and text. Specifcally, we first transform R_i^{vt} and R_i^t into the same modal feature space as follows.

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

[2] https://huggingface.co/bert-base-uncased

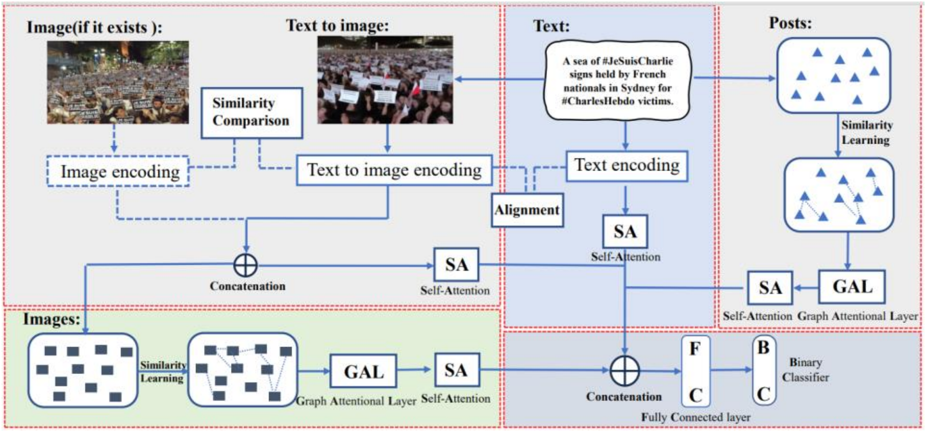


Fig. 2. The architecture of the proposed unified model.

$$R_i^{t'} = W_t R_i^t, \quad R_i^{vt'} = W_{vt} R_i^{vt} \tag{4}$$

where W_t and W_{vt} are learnable parameters. Then we narrow the distance between $R_i^{t'}$ and $R_i^{vt'}$ by the MSE loss for modal alignment is as follows.

$$\mathcal{L}_{align} = \frac{1}{n}\sum_{i=1}^{n}\left(R_i^{t'} - R_i^{vt'}\right)^2 \tag{5}$$

Similarity comparison. After aligning R_i^{vt} and R_i^t through modal alignment, then we can get the similarity between v_i and v_{ti} by calculating the cosine values of R_i^v and R_i^{vt} as follows.

$$\alpha = (R_i^v * R_i^{vt})/\left(||R_i^v||||R_i^{vt}||\right) \tag{6}$$

If the similarity α is less than 0.5, we will think that the image v_i and text t_i are inconsistent and remove v_i. Otherwise, we will concatenate the R_i^v and R_i^{vt} for a new R_i^v as follows.

$$R_i^v = concat(R_i^v, R_i^{vt}) \tag{7}$$

Since the text graph structure and the image graph structure are processed similarly, next we will explain the image graph structure specifically. We first use Eq. (6) to similarly calculate the similarity between images and texts in posts. If the similarity is higher than 0.7, we will place an edge between them. The formula is shown in Eq. (8), where e_{ij} stands for whether existing an edge between the features of images R_i^v and R_j^v, $e_{ij} = e_{ji}$.

$$e_{ij} = \begin{cases} 1, & if\ \alpha_{ij} > 0.7 \\ 0, & otherwise \end{cases} \tag{8}$$

Graph attentional layer. The next we can obtain the similarity information through Graph Attentional Layer (GAL). The key of GAL is the aggregation of the neighborhood information. For node n_i, we first get its neighbor nodes $\mathcal{N}_i = \{n_i^1, n_i^2, n_i^3, \ldots, n_i^j\}$, where j is the number of neighbor nodes and n_i^j is the neighbor node. We first calculate the attention weight $\beta = \{e_i^1, e_i^2, e_i^3, \ldots, e_i^j\}$ between n_i and each node in $\mathcal{N}_i$, the formula is shown in Eq. (9), where $\oplus$ denotes concatenation of vectors, κ and W are learnable parameters, x_i and x_j' are node embeddings of n_i and its neighbor nodes n_i^j in $\mathcal{N}_i$.

$$e_i^j = LeakyReLU\left(\kappa\left[Wx_i \oplus Wx_j^{'}\right]\right) \tag{9}$$

Then, we use the softmax function to perform weight normalization on the attention weights. After that, the normalized attention coefficients are used to compute a linear combination of the features corresponding to them to serve as the final output features for every node. Finally, a multi-head attention mechanism [23] is adopted to capture features from different perspectives. The formula is shown in Eq. (10), where e_i^j is the attention weight in β, R_i^{gv} is the graph feature of images, H denotes the number of heads, x_i^j is the embedding of the node in $\mathcal{N}_i$, $\oplus$ denotes concatenation of vectors. Similarly, we can get the graph feature of texts R_i^{gt}.

$$R_i^{gv} = \bigoplus_{h=1}^{H} \sigma\left(\sum_{j\in\mathcal{N}_i} softmax_i\left(e_i^j\right)^h W^h x_i^j\right) \tag{10}$$

where σ is a nonlinear activation function.

Self-attention. Next, we use the self-attention [23] to enhance the features of R_i^t, R_i^v, R_i^{gv}, and R_i^{gt} respectively. Specifically, We use the following equation to calculate the query matrix, key matrix and value matrix, respectively, where R_i^t is taken as an example, $W^Q, W^K, W^V \in R^{d\times\frac{d}{H}}$ are linear transformations:

$$Q_i^t = R_i^t W^Q, \quad K_i^t = R_i^t W^K, \quad V_i^t = R_i^t W^V \tag{11}$$

Then we can get the more representative text features $R_i^{t''}$, and the formula is shown in Eq. (12), where H denotes the number of heads, $\oplus$ denotes concatenation of vectors, and $W_t^O \in R^{d\times d}$ is the output linear transformations.

$$R_i^{t''} = \left(\bigoplus_{h=1}^{H} softmax\left(\frac{Q_i^t K_i^t}{\sqrt{d}}\right)V_i^t\right)W_t^O \tag{12}$$

Similarly, we can get $R_i^{v''}$, $R_i^{gv''}$, $R_i^{gt''}$.

Fully connected layer. Finally, we concatenate and feed $R_i^{v''}$, $R_i^{gv''}$, $R_i^{gt''}$ into the fully connected layer to predict whether p_i is a rumor or not:

$$\widehat{y_i} = softmax\left(W_r concat\left(R_i^{t''}, R_i^{v''}, R_i^{gv''}, R_i^{gt''}\right) + b\right) \tag{13}$$

Table 1. The statistics of two datasets.

Statistics	tweets	images	non-rumors	rumors
PHEME	5746	2018	3653	2093
Weibo	4664	3842	2351	2313

where W_r and b are the trainable weight matrix and bias, respectively, $\hat{y}$ is the final prediction result.

Objective function. The rumor detector is trained with cross-entropy loss against the ground-truth distribution y_i, and the formulas for classification loss and total loss are:

$$\mathcal{L}_{classify} = -y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i) \tag{14}$$

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

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

4 Experimentation

4.1 Datasets

We evaluate our model on two real-world datasets: Weibo [9] and PHEME [24]. The language of the Weibo dataset is Chinese and is collected from Weibo, one of the most popular social platforms in China. The language of the PHEME dataset is English, collected from Twitter, and its main content is 5 breaking news. Since some baseline models need to contain both text and images, we experimented GAGIN on the datasets that contain both text and images and the entire dataset. The statistical results of the two datasets obtained after removal are shown in the **Table 1**.

4.2 Experimental settings

For both datasets, we use similar preprocessing methods: 1). Removing the URL part of the text. 2). Removing data containing only plain URLs or plain "@xxx". 3). Each tweet extracted up five comments at most. 4). Images were resized to 224 × 224 pixels and normalized.

We use BERT to initialize word embeddings of size 768. We use Adam [25] to optimize our objective function. The number of heads H is set to 6. λ_a and λ_c are set to 1.6 and 2.2. For the fair comparison, we perform 5-fold cross-validation in all experiments and report average results.

4.3 Baselines

We compare the GAGIN model to the baselines listed below.

- ✧ **Text-CNN** [26] is a deep learning model designed for text classification tasks.

Table 2. The results of GAGIN and baselines on PHEME.

Method	PHEME			
	Acc. (%)	**Pre. (%)**	**Rec. (%)**	**F1**
Text-CNN	63.6	40.4	63.6	49.4
BERT	85.4	84.1	84.5	84.3
EANN	78.4	74.5	77.3	95.9
MVAE	83.1	84.1	83.1	83.4
MFAN	87.4	87.7	87.4	87.5
MGIN-AG	87.5	84.4	86.8	85.4
GAGIN/m	87.8	86.9	86.5	86.7
GAGIN	**88.5**	**87.8**	**87.3**	**87.5**

- ✧ **BERT** [11] is currently the most popular pretrained language representation model.
- ✧ **EANN** [3] is a multimodal model where VGG-19 encoded image features and w2v encoded text features.
- ✧ **MVAE** [2] is a multimodal variational auto-encoder that can effectively learn shared representations between images and text.
- ✧ **MFAN** [20] is multimodal feature-enhanced attention network based on self-attention.
- ✧ **MGIN-AG** [21] is interactive network between the words of the text and image blocks.

Among them, Text-CNN and BERT are unimodal methods, while EANN, MVAE, MFAN and MGIN-AG are all multimodal ones, most of which use Text-CNN for text encoding. The multimodal methods and GAGIN/m choose datasets that contains both text and image parts, while the others select the entire dataset.

4.4 Results

Table 2 and **Table 3** shows the performance of all methods, and the results show that our GAGIN model outperforms all baselines and we also draw the following observations. Analysis can be conducted according to the following aspects:

1) Most of them are based on text-CNN methods to learn text features. That is, sentence features are trained by training word vectors, and word vectors are trained by training dictionaries, without considering the position information of the word in the sentence. However, BERT, as a pre-training model, takes these into consideration, so most of these models do not perform as well as directly using BERT for single text encoding training.

2) MVAE is improved on basis of EANN, so the effect is obviously better than EANN. On basis of already having image and text information, MFAN adds an additional social graph structure, so its performance is better than EANN and MVAE.

Table 3. The results of GAGIN and baselines on Weibo.

Method	Weibo			
	Acc. (%)	**Pre. (%)**	**Rec. (%)**	**F1**
Text-CNN	74.3	83.1	74.1	72.3
BERT	89.1	89.1	89.1	89.1
EANN	80.7	83.0	80.7	81.8
MVAE	85.2	85.5	85.3	85.4
MFAN	90.1	90.1	90.1	90.1
MGIN-AG	93.3	93.4	93.2	93.3
GAGIN/m	93.8	93.7	93.8	93.8
GAGIN	**94.6**	**94.7**	**94.6**	**94.6**

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

Method	Weibo		PHEME	
	Acc. (%)	**F1**	**Acc. (%)**	**F1**
GAGIN	**94.6**	**94.6**	**88.5**	**87.5**
w/o IG	90.3	90.3	84.5	82.0
w/o SC	92.8	92.8	86.7	86.6
w/o A	93.4	93.2	87.8	87.1
w/o G	94.0	94.0	88.3	87.3

3) MGIN-AG only considers representations of the same modal features. However, it did not take the connections into account and inconsistencies between different modalities, so the effect is not as good as our GAGIN/m. Since the data of GAGIN is more complete than the training sample data of GAGIN/m, GAGIN is slightly better than GAGIN/m.

4.5 Ablation study

To verify the effectiveness of each module of GAGIN, we consider the following variants by removing one of the components in the model:

- **w/o** IG: Removing the images generated according to texts from GAGIN.
- **w/o** SC: Removing the similarity comparison between images.
- **w/o** A: Removing alignment between images and texts.

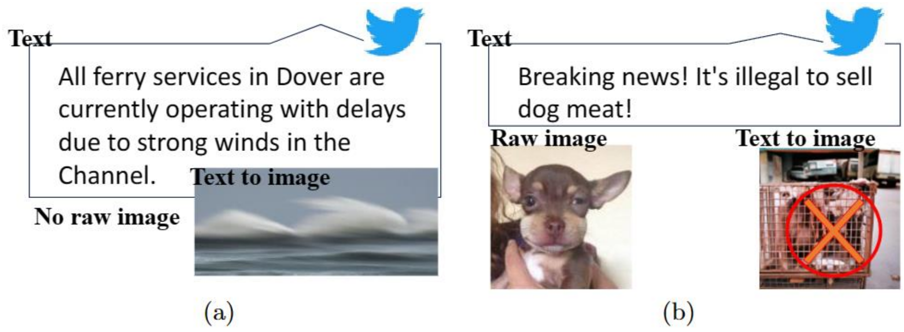


Fig. 3. Two typical cases detected by GAGIN, where (a) a sample with only texts where the image is generated; (b) a sample with the text and image but they are inconsistent.

- **w/o** G: Removing graphical information between texts or images.

The experimental results are shown in **Table 4** and we can draw the following observations.

1) The simplified model w/o IG achieves the relatively lowest results. The reason is that it not only fails to avoid the interference of inconsistent images and text, but also fails to make the model better understand based on the generated images, causing the subsequent butterfly effect. This result proved the effectiveness of our mechanism of generating images for pure text post.
2) Compared with GAGIN, w/o SC has a significant decrease on accuracy (Weibo/PHENE: -1.8/-1.8). This result shows that similarity comparison can solve the problem of the inconsistency between images and texts to a certain extent.
3) The model w/o A performs worse than GAGIN. Modal alignment is to enable different modes of similar things to form similar expressions in space. Removing the module modal alignment will result in no more relevant and representative feature representation being obtained between texts and images.
4) The model w/o G has the relatively lowest impact but it's also important. The graph structure is equivalent to a guarantee. When text or image information is poorly learned, the graph structure can play a corrective role at this time.

4.6 Case study

To further illustrate the effectiveness of our GAGIN, we give two representative cases, all of which have been successfully classified by our model.

It can be seen that, in **Fig. 3(a)**, for plain text, the lack of image is more difficult for the detector to understand than having both image and text. Therefore, we added image information that matches the text so that the detector can better comprehend the tweet.

In **Fig. 3(b)**, inconsistencies between images and text may cause the detector to understand unnecessary information. Hence, we compare the similarity of the images, and then remove the raw image and use the generated image that is more consistent with the text, allowing the model to detect the rumor without interference.

5 Conclusion

In this paper, we propose a novel unified model, namely Graph Attention Generative Image Network (GAGIN), which can be used for unimodal or multimodal rumor detection. Specifically, we generate images based on corresponding texts, so that our multimodal model can be applied to text-only tasks and be benefit from the interaction between texts and images. Through our similarity comparison and modal alignment mechanisms, more significant image and text features can be obtained. Experimental results on English Pheme and Chinese Weibo show that our GAGIN outperforms the state-of-the-art baselines. Our future work will focus on how to select the highly correlated images from the set of generated and posted images for multimodal rumor detection.

Acknowledgments. The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Key R&D Plan of Jiangsu Province (BE2021048).

References

1. Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, et al., "Covid-19–related infodemic and its impact on public health: A global social media analysis," The American journal of tropical medicine and hygiene, vol. 103, no. 4, pp. 1621, 2020.
2. Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in WWW, 2019, pp. 2915–2921.
3. Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, et al., "Eann: Event adversarial neural networks for multi-modal fake news detection," in ACM SIGKDD, 2018, pp. 849–857.
4. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in CVPR, June 2022, pp. 10684–10695.
5. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," 2021.
6. Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., "Graph attention networks," stat, vol. 1050, no. 20, pp. 10–48550, 2017.
7. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, "Information credibility on twitter," in WWW, 2011, pp. 675–684.
8. Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang, "Automatic detection of rumor on sina weibo," in ACM SIGKDD, 2012, pp. 1–7.

9. 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,” in IJCAI. 2016, AAAI Press.
10. Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al., “A convolutional approach for misinformation identification.,” in IJCAI, 2017, pp. 3901–3907.
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
12. Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan, “Kan: Knowledge-aware attention network for fake news detection,” in AAAI, 2021, vol. 35, pp. 81–89.
13. Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun, “Different absorption from the same sharing: Sifted multi-task learning for fake news detection,” arXiv preprint arXiv:1909.01720, 2019.
14. Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu, “Mining dual emotion for fake news detection,” in WWW 2021, 2021, pp. 3465–3476.
15. Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu, “defend: Explainable fake news detection,” in ACM SIGKDD, 2019, pp. 395–405.
16. Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu, “Hierarchical propagation networks for fake news detection: Investigation and exploitation,” in AAAI, 2020, vol. 14, pp. 626–637.
17. Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
18. Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu, “Hierarchical multi-modal contextual attention network for fake news detection,” in ACM SIGIR, 2021, pp. 153–162.
19. Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu, “Multimodal fusion with co-attention networks for fake news detection,” in ACL, 2021, pp. 2560–2569.
20. Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang, “Mfan: Multi-modal feature-enhanced attention networks for rumor detection,” in IJCAI, 2022.
21. Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu, “Graph interactive network with adaptive gradient for multi-modal rumor detection,” in ICMR, 2023, pp. 316–324.
22. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.
23. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” NeurIPS, vol. 30, 2017.
24. Arkaitz Zubiaga, Maria Liakata, and Rob Procter, “Exploiting context for rumour detection in social media,” in SocInfo 2017, 2017. Springer, 2017, pp. 109–123.
25. Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
26. Yoon Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.