Reinforced Multi-modal Circulant Fusion based Transformers for Rumor Detection

1st Xingang Wang
School of Computer Science and Technology
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
xgwang@qlu.edu.cn

2nd Xiaomin Li
School of Computer Science and Technology
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
xiaominliqlu@163.com

3rd Xiaoyu Liu
School of Computer Science and Technology
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
XiaoyuLiuQlu@163.com

***Abstract*—While the popularity of social media platforms facilitates people's communication, it also contributes to the diffusion of rumors, which may endanger social and public order. Multi-modal rumors in the form of text and images are more likely to gain public trust than traditional text-only content. As a result, the rapid identification of multi-modal rumor information in social media has emerged as a significant research focus today.**

Although multi-modal rumor detection has been widely researched, we still face a challenge. We found that the current multi-modal fusion methods for this task mainly include element-wise product, element-wise sum, or even direct splicing, which may not better exploit the advantages of the complementary nature of multi-modal data. Moreover, previous studies used the fused features immediately for downstream tasks, which lacks further exploration of fused featured.

To solve the above issues, we propose Reinforced Multi-modal Circulant Fusion(MCF) based Transformers for Rumor Detection (MCFbT), which considers further optimization of the fused multi-modal features. Firstly, we use two pre-trained models, ALBERT and VGG-19, to extract the textual and visual features of the posts. Secondly, we fuse the two features using MCF to obtain the initial multi-modal features. Thirdly, we enhance the multi-modal features via the cross-modal transformers. Finally, the optimized multi-modal features are fed into the detector to identify rumors. We performed extensive experiments on two public datasets to compare previous models, and the results show that the model outperforms the baseline models.

***Index Terms*—Social Media, Rumor Detection, Multi-modal Learning, Multi-modal Circulant Fusion, Cross-modal Transformer**

I. Introduction

With the continued development of society and the advancement of science and technology, social media has become the primary way for people to obtain and exchange information. Still, it has also accelerated the breeding and rapid spread of rumors. Rumors frequently confuse and mislead the masses, causing social panic and disrupting social harmony and stability. Compared with plain text, posts with images are more likely to deepen users' impressions, increase the credibility of information, and mislead users to retweet and spread it. Rumor publishers began using extremely misleading and even doctored images to attract readers' attention. As a result, in recent years, multi-modal rumor detection based on social media has become a major research topic.

Although some progress has been made in developing multi-modal rumor detection models, they still face a major problem. Visual and text features are in different semantic feature spaces, and there is heterogeneity. Current fusion methods include directly splicing, element-wise product, and element-wise sum. So simply fusing features lacks in-depth data analysis and ignores complex local interconnections, which obviously cannot better utilize the complementary information of multi-modal data. Moreover, we are concerned that previous studies [20] have mainly adhered to the fixed idea of detection immediately after obtaining multi-modal features, which we believe may hinder accuracy improvement.

To solve the above issue, we propose Reinforced Multi-modal Circulant Fusion based Rumor Detection Transformers (MCFbT), which adds an optimized multi-modal feature mechanism after fusion. Firstly, We extract the textual and visual features of posts using the pre-training models ALBERT and VGG-19, respectively. Secondly, using the MCF [25] method fuses textual and visual features to obtain the initial multi-modal features. Thirdly, the multi-modal features are reinforced using cross-modal transformers [21] from the two modalities. Finally, the final features are used to distinguish rumors.

Overall, in this paper, we make the following contributions:

- In the multi-modal rumor detection task, We consider for the first time allowing the individual modalities to enrich the multi-modal features by cross-modal transformers, and the optimized multi-modal features focus more on the key parts of the information.
- We evaluate MCFbT on two public real-world datasets. The results demonstrate that the model outperforms previous models.

II. Related Work

In this section, previous research is briefly described in terms of both text-based rumor detection and multi-modal rumor detection.

A. Text-based Rumor Detection

Existing methods [4], [14], [15] focus mainly on extracting features from the text content of posts, which has been extensively researched in many papers on rumor detection. For

This work is supported in part by The 20 Planned Projects in Jinan(202228120), Natural Science Foundation Project of Shandong Province(ZR2022LZH008)

DOI reference number:10.18293/SEKE23-162

instance, Castillo et al. [4] investigate behavioral characteristics and conclude that users who have posted more messages are more likely to post truthfulness based on the posters' previous posting history. Ma et al. [13] used TF-IDF to represent news texts and then used a recurrent neural network(RNN) to model the Weibo dataset. Chen et al. [5] proposed an RNN-based deep attention model to identify rumors by selectively learning the temporal representation of sequential posts and validated the effectiveness of RNN in the study. Potthast et al. [15] conducted several experiments using the random forest algorithm to compare the effectiveness of topic-based posts and style-based posts features for identifying rumors and found that style-based analysis only did not identify rumors. Yu et al. [26] attempted to use CNN for early rumor detection by extracting multi-scale text features through N-gram windows of varying sizes. Liao et al. [12] segmented tweets into different intervals and used a two-layer GRU network with an attention mechanism to obtain the potential features of tweets and temporal sequences.

B. Multi-modal Rumor Detection

It has been shown that images can provide effective visual information. Therefore, it is necessary to consider research based on multimodal forms in rumor detection, i.e., using text and image information to detect rumors.

In the multimodal rumor detection, Jin et al. [10] introduced neural networks to fuse different modality features of posts and attention mechanism to extract contextual relevant information for rumor detection. Wang et al. [24] combined the idea of adversarial learning to propose the EANN model, which adds the sub-task of event classification to the original classification task and induces the model to learn event-independent features to improve the generalization ability of the model. Khattar et al. [11] proposed an end-to-end multi-modal variational self-encoder model (MVAE) that learns the potential distribution of multi-modal features to explore the distribution pattern of rumor. Chen et al. [7] proposed the CAFE rumor detection model, which can adaptively aggregate unimodal and cross-modal correlations and thus improve task accuracy. Qi et al. [16] innovatively extract visual entities (e.g., celebrities and landmarks) to understand the high-level semantics of post-related images and then model the inconsistency and mutual enhancement of multi-modal entities with the help of visual entities. Dhawan et al. [9] proposed GAME-ON, an end-to-end trainable framework based on a graph neural network that allows granularity interactions within and across different modalities to learn more robust data.

Unlike all the above work, we consider feature enhancement after the fusion of images and text when identifying rumors on social media.

III. Methodology

In this section, we mainly introduce the proposed model MCFbT in detail. As shown in Fig. 1, MCFbT has four parts:

- Multi-modal Feature Extractor. a pre-trained model ALBERT, is used to model the text and obtain the text features, while the VGG-19 network is used to extract the visual feature of images.
- Multi-modal Feature Fusion. The MCF method is used to fuse visual and textual features to obtain the initial multi-modal features.
- Multi-modal Feature Enhancement. To obtain a more expressive feature representation, we use an architecture consisting mainly of cross-modal converters to optimize the initialized multimodal features.

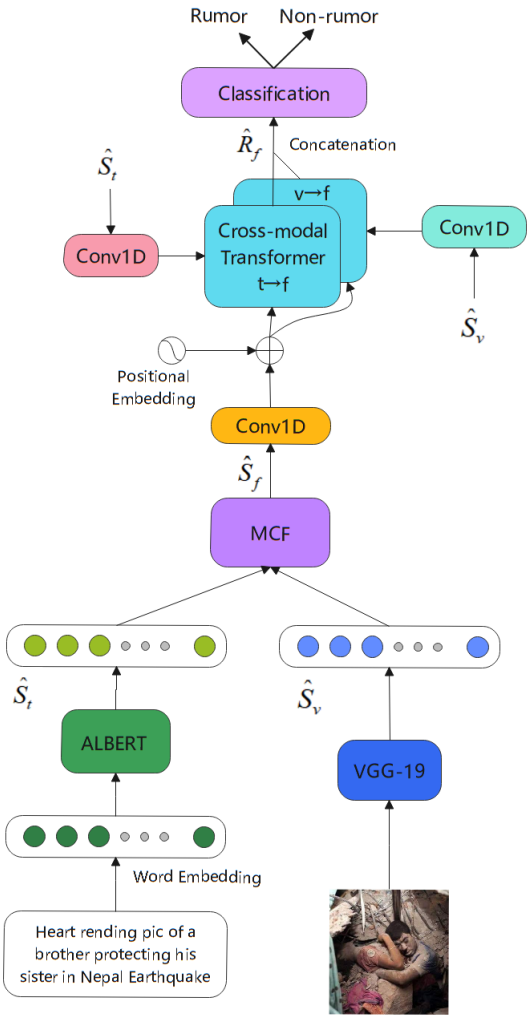

Fig. 1. The overall framework of MCFbT.

A. Multi-Modal Feature Extractor

To accurately model the semantic and contextual meaning of words, the paper uses the ALBERT [8] model with 12 encoder layers for text feature extraction. ALBERT ensures the quality of feature extraction while reducing parameters and improving training speed. The model is used to obtain features of the complete text, $S_t \in \mathbb{R}^{1\times d_t}$, where d_t denotes the dimensionality of the text features obtained from the ALBERT, we do not fine-tune the ALBERT weights, so the ALBERT network's weights are frozen.

We refer to previous research [17], [22], [27], and choose to use the VGG-19 [19] network to extract visual features from images.This is a 19-layer VGGNet pre-trained on the ImageNet dataset [19] to obtain visual feature vectors, $S_v \in \mathbb{R}^{1\times d_v}$ where d_v is the visual feature dimension output by VGG-19. The pre-trained model is fixed during training, and

the weights of VGG-19 are not fine-tuned, and the weights of the VGG network are frozen.

Since it is possible that $d_t \neq d_v$, we design that a fully connected layer is added to ensure the same dimensionality d as the two features, as follows:

$$\hat{S}_t = \mathrm{ReLU}\left(S_t \times W_t + b_t\right) \tag{1}$$

$$\hat{S}_v = \mathrm{ReLU}\left(S_v \times W_v + b_v\right) \tag{2}$$

where $W_t \in \mathbb{R}^{d_t \times d}$ and $W_v \in \mathbb{R}^{d_v \times d}$ are the weight matrices, and b_t and b_v are the bias terms.

B. Multi-modal Feature Fusion

The paper adopts the MCF method to achieve feature fusion. The method uses the newly defined interaction operation to complete the fusion after transforming the feature vector into circulant matrices [25].

We use the projection vector $X \in R^d$ and $Y \in R^d$ to construct circulant matrix $A \in \mathbb{R}^{d \cdot d}$ and $B \in \mathbb{R}^{d \cdot d}$.

$$A = \mathrm{circ}(\hat{S}_t) \tag{3}$$

$$B = \mathrm{circ}(\hat{S}_v) \tag{4}$$

where $circ(d)$ denotes converting d to a circulant matrix.

Then the projection vector and each row vector of the circulant matrix do an element-wise product. The algorithm is as follows:

$$M = \frac{1}{d}\sum_{i=1}^{d} a_i \odot \hat{S}_t \tag{5}$$

$$N = \frac{1}{d}\sum_{i=1}^{d} b_i \odot \hat{S}_v \tag{6}$$

where $a_i \in R^d$ and $b_i \in R^d$ are the row vectors of the circulant matrix A and circulant matrix B respectively, $\odot$ denotes the operation of the element-level product. Crucially, this process does not introduce new parameters.

Finally, through the projection matrix $W_3 \in R^{d \cdot d}$, we convert the element-level sum vectors of $M \in R^d$ and $N \in \mathbb{R}^d$ into the target vectors $\hat{S}_f \in R^d$, the multi-modal features.

C. Multi-modal Feature Enhancement

We obtain textual features $\hat{S}_t$, visual features $\hat{S}_v$ and multi-modal features $\hat{S}_f$ through the above two subsections, all of which have dimensions of d. Below we will introduce the multi-modal features optimization architecture based on transformers [18].

Temporal Convolutions. To ensure that each element in the input sequence has a sufficient understanding of its neighbors, we pass the input sequence through a 1D temporal convolutional layer:

$$R_{\{t,v,f\}} = \mathrm{Conv1D}\left(\hat{S}_{\{t,v,f\}}, k_{\{t,v,f\}}\right) \in \mathbb{R}^{T}{}_{\{t,v,f\}} \times d \tag{7}$$

where $k_{(t,v,f)}$ are the sizes of the convolutional kernels for modalities (t, v, f), and d is a common dimension.

Positional Embedding. In order to enable multi-modal feature sequences to carry temporal information, We add positional embedding (PE) to R_f:

$$\hat{R}^{[0]}_{t,v,f} = R_{t,v,f} + PE\left(T_{t,v,f}, d\right) \tag{8}$$

where $PE\left(T_{t,v,f}, d\right) \in \mathbb{R}^{T_{t,v,f} \times d}$ calculate the (fixed) embedding vectors for each position index, and $\hat{R}^{[0]}_{t,v,f}$ is a low-level location-aware feature generated for multi-modal features.

Cross-modal Transformers. We refer to the cross-modal transformer designed in previous studies [21], which enables one modality to receive information from another modality. We assume that the textual information (t) is passed to the multi-modal information (f), denoted by "$t \rightarrow f$". The complete calculation process is as follows

$$\hat{R}^{[0]}_{t\rightarrow f} = Z^{[0]}_f \tag{9}$$

$$\hat{R}'^{[i]}_{r\rightarrow f} = \mathrm{CM}^{[i],\mathrm{mul}}_{t\rightarrow f}\left(\mathrm{LN}\left(\hat{R}^{[i-1]}_{t\rightarrow f}\right), \mathrm{LN}\left(\hat{R}^{[0]}_{t}\right)\right) + \mathrm{LN}\left(\hat{R}^{[i-1]}_{t\rightarrow f}\right) \tag{10}$$

$$\hat{R}^{[i]}_{t\rightarrow f} = f_{\theta^{[i]}_{t\rightarrow f}}\left(\mathrm{LN}\left(\hat{R}'^{[i]}_{r\rightarrow f}\right)\right) + \mathrm{LN}\left(\hat{R}'^{[i]}_{r\rightarrow f}\right) \tag{11}$$

where f_θ is a positionwise feed-forward sublayer parametrized by θ, and $\mathrm{CM}^{[i],\mathrm{mul}}_{t\rightarrow f}$ means a multi-head version of $\mathrm{CM}_{t\rightarrow f}$ [21] at layer i. LN means layer normalization [2].

$\hat{R}^{[i]}_{v\rightarrow f}$ follows the same steps as described above. Then, we concatenate the outputs from the cross-modal transformers to obtain the final multi-modal features $\hat{R}_f = \left[\hat{R}^{[D]}_{t\rightarrow f}; \hat{R}^{[D]}_{v\rightarrow f}\right]$, $\hat{R}_f \in \mathbb{R}^{T_f \times 2d}$.

D. Rumor Detection

We project the multi-modal feature vector $\hat{R}_f$ into the target space of whether it is a rumor or not using a fully connected layer with softmax activation and obtain the probability distribution p:

$$p = \mathrm{softmax}\left(W\hat{R}_f + b\right) \tag{12}$$

where W is the weight matrix and b is the bias term.

To calculate the classification loss, we use the cross-entropy loss as follows:

$$\mathcal{L}_p = -\left[y \log p_0 + (1-y)\log p_1\right] \tag{13}$$

where $y \in \{0, 1\}$ denotes the ground-truth label.

IV. Experiments

A. Datasets

To assess the effectiveness of the MCFbT, we conducted experiments on two real-world datasets, which are collected from Twitter and Weibo, respectively. The Twitter dataset was released for Verifying Multimedia Use task at MediaEval [3]. It is divided into two parts, the development set and the test set, with a ratio of 7:3, and there is no overlap between them. We use the development set for training and the test set for evaluation to compare baselines fairly. The Weibo dataset is collected by Jin et al. [10], basic situation of which is similar to the Twitter. The two datasets use images and text, so we remove posts without any text or images. Table II shows the statistics of the two datasets.

TABLE I
THE RESULTS OF DIFFERENT METHODS ON TWO DATASETS

Dataset	*Method*	*Accuracy*	Rumour			Non-rumour		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	Text-Only	0.706	0.648	0.540	0.589	0.715	0.636	0.673
	Image-only	0.596	0.695	0.518	0.593	0.524	0.700	0.599
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.650
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.682	0.780	0.615	0.689	0.603	0.770	0.676
	EANN	0.719	0.642	0.474	0.545	0.771	**0.870**	**0.817**
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	CAFE	0.806	0.807	0.799	0.803	0.05	0.813	0.809
	MCFbT	**0.842**	**0.830**	**0.863**	**0.846**	**0.850**	0.735	0.777
Weibo	Text-Only	0.804	0.800	0.860	0.830	0.840	0.760	0.800
	Image-only	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	NeuralTalk	0.726	0.794	0.613	0.692	0.684	0.840	0.754
	att-RNN	0.788	**0.862**	0.686	0.764	0.738	**0.890**	0.807
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCFbT	**0.870**	0.856	**0.875**	**0.865**	**0.859**	0.861	**0.860**

TABLE II
THE STATISTICS OF THE REAL-WORLD DATASETS.

Dataset	Label	Number	all
Twitter	fake	7021	12995
	real	5974	
Weibo	fake	4749	9528
	real	4779	

B. Baseline Model

To verify the validity of this model, we selected the baseline model from the following two models: uni-modal models, multi-modal models.

(i) Uni-modal models

- Text-only. A model that uses only text information and extracts textual features by ALBERT.
- Image-Only. A model that uses only image information and extracts visual features by VGG-19.

(ii) Multi-modal models

- VQA. VQA [1] is a model for answering questions about a given image. For comparison with MCFbT, the VQA is designed for a binary classification task.
- NeuralTalk. NeuralTalk [23] is a deep recurrent framework for image caption. To adapt rumor detection task, the model feature is averaged over the RNN output at each time step.
- att-RNN. The att-RNN [10] uses an attention mechanism to combine textual, visual and social context features. In the paper, the part dealing with social environment information is removed from the experiments.
- EANN. EANN is the Event Adversarial Neural Networks proposed in [24]. The framework is to guide the model to learn event-independent multi-modal features by introducing an event classifier as a secondary task.
- MVAE. MVAE [11] is a multi-modal model based on Variational Autoencoder. The spliced features are encoded as an intermediate expression for reconstructing input features and rumor classification.
- CAFE. CAFE [6] is a multi-modal rumor detection method with fuzzy perception. The model can wisely and adaptively aggregate uni-modal features and cross-modal correlations.

C. Results and Analysis

Table I shows the results of baselines and our proposed model on two datasets. The experimental results show that our proposed method outperforms the baseline. The following are some specific observations:

- Text-based method outperforms images-based approach, proving that text contains more information than images.
- The multi-modal models outperform the uni-modal model, showing that multi-modal methods generally outperform uni-modal based methods, demonstrating the superiority of multi-modal Features.
- Among the multi-modal models tested in the baseline, MCFbT performed the best, indicating that the model can accurately capture more effective multi-modal features to detect rumors.

To demonstrate the importance of each component, we perform ablation analysis on the MCFbT model and experiment on Twitter and Weibo datasets, the following variants of our model are designed for comparison:

- Base: the most basic model composed of the ALBERT, the VGG-19 and the MCF.
- Base-tf: models augmented with multi-modal features rely only on textual features.

- Base-vf: models augmented with multi-modal features rely only on visual features.
- MCFbT: The whole model MCFbT with all components.

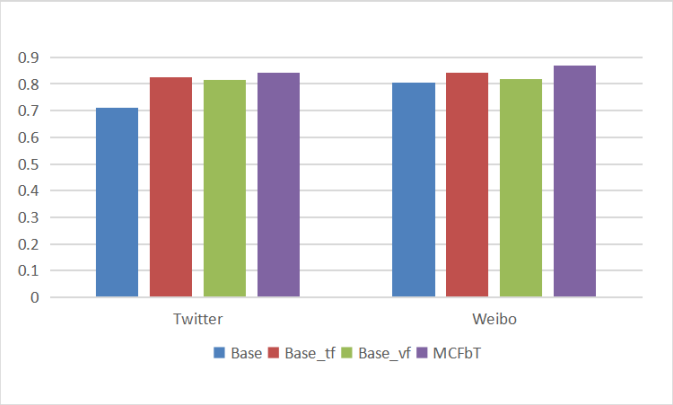

Fig. 2. The performance comparison for variants of the model.

As shown in Fig. 2, both Base-tf and Base-vf are more effective than the Base, and the overall model MCFbT results are higher than the Base-tf and the Base-vf, indicating that each component is necessary and effective. In particular, we find that Base-tf is better than Base-vf on both datasets, suggesting that textual features cover richer semantic information to help us detect rumors.

V. CONCLUSION

In this paper, we propose MCFbT, Reinforced Multi-modal Circulant Fusion based Transformers for Rumor Detection. In this task, we consider using textual and visual features to optimize multi-modal features after fusion via cross-modal Transformers. Compared with the traditional method, the model makes it pays more attention to vital information and relatively reduces the influence of noise. We conduct extensive experiments on two real-world datasets, the results demonstrate the effectiveness of our proposed method. In the future, we plan to conduct research from a visual information perspective to contribute to rumor detection with potentially better results.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3):7, 2015.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[5] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 40–52. Springer, 2018.

[6] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2897–2905. ACM, 2022.

[7] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, page 2897–2905, 2022.

[8] 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.

[9] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478*, 2022.

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

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

[12] XW Liao, Zhi Huang, DD Yang, Xueqi Cheng, and Guolong Chen. Rumor detection in social media based on hierarchical attention network. *Sci Sin Inform*, 48(11):1558–1574, 2018.

[13] 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.

[14] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450, 2012.

[15] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[16] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 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*, page 1212–1220, 2021.

[17] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 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*, pages 1212–1220, 2021.

[18] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman. Low rank fusion based transformers for multimodal sequences. 2020.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437, 2021.

[21] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[22] Nguyen Manh Duc Tuan and Pham Quang Nhat Minh. Multimodal fusion with bert and attention mechanism for fake news detection. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE, 2021.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[24] 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 *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[25] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *IJCAI*, volume 3, page 8, 2018.

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

[27] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951, 2019.