ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747280

CROSS-MODAL KNOWLEDGE DISTILLATION IN MULTI-MODAL FAKE NEWS DETECTION

Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, Dongsheng Li

College of Computer, National University of Defense Technology
{weizimian16, hengyuepan, qiao.linbo, niuxin, dongpeijienudt, dsli}@nudt.edu.cn

ABSTRACT

Since the rapid dissemination of fake news brings a lot of negative effects on real society, automatic fake news detection has attracted increasing attention in recent years. In most circumstances, the fake news detection task is a multi-modal problem that consists of textual and visual contents. Many existing methods simply integrate the textual and visual features as a shared representation but overlook their correlations, which may lead to sub-optimal results. To address this problem, we propose CMC, a two-stage fake news detection method with a novel knowledge distillation that captures Cross-Modal feature Correlations while training. In the first stage of CMC, the textual and visual networks are trained mutually in an ensemble learning paradigm. The proposed cross-modal knowledge distillation function is presented as a soft target to guide the training of a single-modal network with the correlations from the other peer. In the second stage of CMC, the two well-trained networks are fixed, and their extracted features are fed to a fusion mechanism. The fusion model is then trained to further improve the performance of multi-modal fake news detection. Extensive experiments on Weibo, PolitiFact, and GossipCop databases show that CMC outperforms the existing state-of-the-art methods by a large margin.

Index Terms— fake news detection, knowledge distillation, multi-modal

1. INTRODUCTION

Automatic fake news detection is important for normal society to avoid the rampant dissemination of fake news on social media and the Internet. Generally, the main task of fake news detection is to identify fake news according to the extracted features. The features can be obtained from various sources such as textual contents, attached images, social contexts, etc. In this paper, we mainly focus on textual and visual features.

Previous textual-visual-based methods can be divided into two classes: single-modal methods and multi-modal methods. Qi et al. [1] proposed a single-modal method that introduces a multi-branch CNN-RNN model to extract visual features from frequency and pixel domains. [2] construct a ensemble classifier by eight different transformer-based pre-trained models to deal with textual news. Compared to single-model methods, multi-modal methods contain more plentiful information by extracting features from different modalities. [3] used a bi-modal variation auto-encoder to learn a shared representation between the textual and visual networks. Spotfake+ [4] integrated pre-trained language transformers and ImageNet models by multiple fully connected layers. However, the drawback of these multi-modal methods is that they overlook cross-modal correlation knowledge, which may lead to sub-optimal results. As mentioned in [5] and [3], the correlations of representations from different modalities is crucial in multi-modal fake news detection. Moreover, the corresponding signal spaces of textual and visual networks are different, which may bring about a negative impact on performance when learning a shared feature between the textual and visual networks.

Inspired from DML [6] that an ensemble of networks can learn collaboratively and teach each other throughout the training process, we propose a multi-modal fake news detector called CMC to train two single-modal networks mutually. One key difference is that the target of the distillation loss in DML is to mimic the class posterior of each network with other peers, while CMC aims to exploit feature correlations between modalities. Specifically, CMC consists of two stages. In the first stage (mutual training stage), the two single-modal networks are trained mutually in an ensemble learning paradigm. Meanwhile, a novel distillation loss is introduced to capture cross-modal feature correlations. By minimizing the cross-modal distillation loss, the positive pairs will be pulled together while the negative pairs will separate from each other. In the second stage (fusion mechanism training stage), the parameters of the textual and visual networks are fixed, and only a fusion mechanism based on BLOCK [7] is trained to further improve performance by better fitting discriminative information from different modalities.

The main contributions of this work are: (1) We propose a mutual learning strategy in multi-modal fake news detection. Instead of integrating a shared representation between different modal networks, we collaboratively train the textual and

This study was supported by the National Natural Science Foundation of China under Grant No. 62025208.

visual networks to gain higher performance. (2) We introduce a cross-modal distillation objective function as a soft target to lead the single-modal network to learn feature correlations between different modalities. (3) We conduct extensive experiments on three real-world datasets. The experimental results show the superiority and effectiveness of our proposed method.

2. APPROACHES

Generally, the proposed CMC method consists of two stages: the mutual training stage and the fusion mechanism training stage. The schematics of two stages of CMC are depicted in Fig. 1. We will first introduce the cross-modal distillation method in subsection 2.1. Then we depict the details of the mutual learning process in subsection 2.2. Subsequently, the fusion mechanism is presented in subsection 2.3.

2.1. Cross-modal Knowledge Distillation

We denote the textual network as f^T and the visual network as f^V. Assuming that we have two input samples x_i and x_j, their corresponding textual and visual features are represented as T_i, V_i and T_j, V_j . The target of the cross-modal distillation is to pull together positive pairs (e.g., T_i and V_i) while pushing away negative pairs (e.g., T_i and V_j). We use $\mathbf{T}$ and $\mathbf{V}$ to represent the textual and visual features for convenience.

Given an anchor T_i from $\mathbf{T}$, we formulate a binary classification problem to precisely select a single positive atom V_i out of a candidate set $S = \{V_i, V_1, V_2, ..., V_k\}$ that contains k uniformly sampled negative atoms. Specifically, we define a variable D, which decides whether V_j was drawn from the positive distribution ($D = 1$) or negative distribution ($D = 0$). The prior probabilities on D are as follows:

$$p(D=1) = \frac{1}{k+1}, \quad p(D=0) = \frac{k}{k+1}. \tag{1}$$

We denote the negative distribution as p_n and positive distribution as p_m. Referring to NCE [10], we formulate p_n as a uniform distribution over all atoms from $\mathbf{V}$. With N represent the dataset size, we have following class-conditional probability:

$$p_n(V_j|D=0, T_i) = \frac{1}{N}. \tag{2}$$

Since $p_m(V_j|D=1, T_i)$ is unknown, we model it by introducing a scoring function $\mathcal{H}(.)$ that is trained to achieve a high value for positive pairs and low for negative pairs.

$$p_m(V_j|D=1, T_i) = \frac{\mathcal{H}(T_i, V_j)}{Z} = \frac{\exp\left(\frac{\phi_1(T_i)\cdot\phi_2(V_j)}{\|\phi_1(T_i)\|\cdot\|\phi_2(V_j)\|}\cdot\frac{1}{\tau}\right)}{Z}, \tag{3}$$

where $\phi_1(.)$ and $\phi_2(.)$ are 1×1 convolution layers to transfer T_i and V_j to the same dimension. They are updated during the training process. Cosine similarity is applied to measure the compactness of $\phi_1(T_i)$ and $\phi_1(V_j)$. τ is a temperature that adjusts the concentration level, and Z is the normalizing constant.

The posterior probability for $D = 1$ is as follows:

$$\begin{aligned} &P(D=1|V_j, T_i) \\ &= \frac{p(D=1)\, p_m(V_j|D=1, T_i)}{p(D=1)\, p_m(V_j|D=1, T_i) + p(D=0)\, p_n(V_j|D=0, T_i)} \\ &= \frac{p_m(V_j|D=1, T_i)}{p_m(V_j|D=1, T_i) + \frac{k}{N}} = \frac{\mathcal{H}(T_i, V_j)}{\mathcal{H}(T_i, V_j) + \frac{k}{N}}. \end{aligned} \tag{4}$$

The objective of partial cross-modal distillation for the textual network is formulated as follows:

$$\begin{aligned} \mathcal{L}_{V\to T} &= -\mathop{\mathbb{E}}_{V_j\sim p_m(\cdot|T_i)}\left[\log\left(P(D=1|V_j, T_i)\right)\right] \\ &\quad - k\cdot\mathop{\mathbb{E}}_{V_j\sim p_n(\cdot|T_i)}\left[1-\log\left(P(D=1|V_j, T_i)\right)\right] \\ &= -\mathop{\mathbb{E}}_{V_j\sim p_m(\cdot|T_i)}\left[\log\left(\frac{\mathcal{H}(T_i, V_j)}{\mathcal{H}(T_i, V_j)+\frac{k}{N}}\right)\right] \\ &\quad - k\cdot\mathop{\mathbb{E}}_{V_j\sim p_n(\cdot|T_i)}\left[\log\left(1-\frac{\mathcal{H}(T_i, V_j)}{\mathcal{H}(T_i, V_j)+\frac{k}{N}}\right)\right]. \end{aligned} \tag{5}$$

By minimizing $\mathcal{L}_{T\to V}$, the relevance scoring function $\mathcal{H}(T_i, V_j)$ will be enlarged for positive pairs and decreased for negative pairs.

Since we train with a cohort of two networks, the total cross-modal distillation objective function is the summation of $\mathcal{L}_{T\to V}$ and $\mathcal{L}_{V\to T}$ as follows:

$$\mathcal{L}_{distill} = \mathcal{L}_{T\to V} + \mathcal{L}_{V\to T}. \tag{6}$$

Then the overall objective for two networks f^T and f^V can be formulated as:

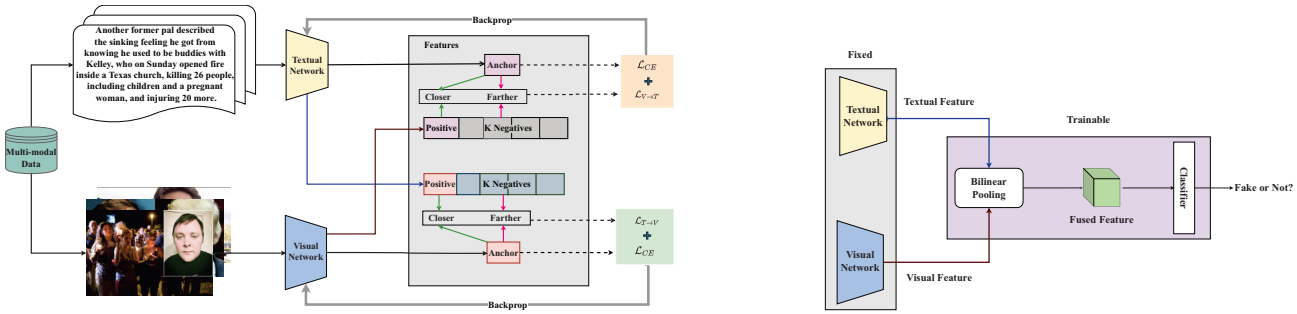
$$\mathcal{L}_{obj_T} = \alpha\cdot\mathcal{L}_{distill} + \mathcal{L}^T_{CE}, \tag{7}$$

$$\mathcal{L}_{obj_V} = \beta\cdot\mathcal{L}_{distill} + \mathcal{L}^V_{CE}, \tag{8}$$

where $\mathcal{L}^T_{CE}$ and $\mathcal{L}^V_{CE}$ are cross-entropy losses of the textual and visual network. α and β are balance parameters to trade off the importance of the cross-entropy loss and cross-modal distillation loss. In practice, we equally set α and β as 0.3.

2.2. Mutual Learning Process

Instead of sharing a concatenated representation, the textual and visual networks perform fake news detection tasks separately. Meanwhile, their training process is closely intervened by each other. The inputs of the two networks are corresponding pre-processed textual and visual contents from the same multi-modal mini-batches. In each iteration, we update the parameters of two networks according to their own predictions and representation correlations with the other peer. The optimizations of the textual and visual networks are conducted iteratively until convergence.



(a) Mutual training (b) Fusion mechanism training

Fig. 1. The pipeline of CMC. In sub-figure (a), the textual network (transformer-based pre-trained model [8],[9]) and the visual network (VGG19 pre-trained on Imagenet) are trained mutually with both the cross-entropy loss ($\mathcal{L}_{CE}$) and the proposed distillation loss ($\mathcal{L}_{V\to T}, \mathcal{L}_{T\to V}$) that transfer cross-modal feature correlations to improve the capacity of each network. In sub-figure (b), the well-trained single-modal networks are fixed, and the fusion mechanism (the bi-linear pooling model [7] and the classifier) is trained to further improve the performance of multi-modal fake news detection.

2.3. Fusion Mechanism

The inputs of the fusion mechanism are extracted features from the fixed textual network and the visual network. Similar to BLOCK [7], the textual feature (x^1) and the visual feature (x^2) are projected to a new feature space by an associate tensor T, specifically:

$$r = T \times_1 x^1 \times_2 x^2, \tag{9}$$

where $\times_1$ and $\times_2$ means tensor product along different dimensional spaces. The final fused tensor r is feed into a SoftMax function to identify fake news.

3. EXPERIMENTATION RESULTS

3.1. Dataset

The Weibo dataset is collected from a social network that is widely used in China called Weibo. As depicted in att-RNN [11], the fake news in Weibo is crawled from an official rumor debunking system ranging from May 2012 to January 2016, and Xinhua News Agency verifies the real news in Weibo. We follow the same method to split the dataset for training and testing as att-RNN [11]. The Politifact and the Gossipcop datasets are collected from the political and entertainment domains of The FakeNewsNet [12] repository, respectively. Each news contains a full-length article and an associated image. According to FakeNewsNet, news of both datasets is checked by domain experts to guarantee the label's credibility. We conduct the same dataset pre-processing method as in Spotfake+ [4] for a fair comparison. The statistics of the three datasets are shown in Table 1.

3.2. Comparison with the state-of-the-art methods

In this section, we show the performances of state-of-the-art methods and CMC in Weibo, PolitiFact, and GossipCop datasets. We used four performance measurements to measure each method, including Accuracy, Precision, Recall, and F1 score. '-' means the results are not available from the original paper. The comparison results are presented in Table 2 .

Table 1. Statistics of datasets

Statistic	Training Set		Test Set		All
	fake	real	fake	real	
Weibo	3749	3783	1000	996	9528
PolitiFact	135	246	29	75	485
GossipCop	2036	7974	545	2285	12840

On the Weibo dataset, Spotfake [13] achieves the best results on Recall and F1 score of fake news. However, CMC outperforms Spotfake at all remaining five metrics, which shows the comparable performance of CMC on Weibo. On PolitiFact dataset, CMC surpasses 2% over the best accuracy achieved by other methods. Moreover, our proposed method exceeds state-of-the-art methods at all metrics on the GossipCop dataset, which further shown the superiority of CMC.

3.3. Ablation Study

We compare CMC with five variants on Weibo to validate the effectiveness of cross-modal knowledge distillation and fusion mechanism. Specifically, we implement five variants as follows:

Finetune-V and Finetune-T: Finetune-V and Finetune-T are single-modal networks in CMC but are trained with a single cross-entropy loss. The inputs of the Finetune-V and Finetune-T networks are single-modal contents that detached from the original multi-modal data.

CMC-V and CMC-T: CMC-V and CMC-T are single-modal networks of CMC that are trained mutually with both a cross-entropy loss and the proposed cross-modal distillation loss.

Table 2. Comparison with other methods on three datasets

Dataset	Method	Acc	Fake News			Real News		
			Prec	Rec	F1	Prec	Rec	F1
Weibo	att-RNN[11]	0.788	0.862	0.686	0.764	0.738	0.89	0.807
	EANN[14]	0.827	0.847	0.812	0.829	-	-	-
	MVAE[3]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	Spotfake[13]	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	MVNN [1]	0.846	0.809	0.857	0.832	-	-	-
	CARMN [15]	0.869	0.935	0.796	0.860	0.820	0.944	0.878
	CMC	**0.908**	**0.940**	0.869	0.899	**0.876**	**0.945**	**0.907**
Politi	RoBERTa-MWSS [16]	0.82	-	-	-	0.82	-	-
	SAFE[5]	0.874	-	-	-	0.889	0.903	0.896
	Spotfake+[4]	0.846	-	-	-	-	-	-
	TM [17]	0.871	-	-	-	0.901	-	-
	LSTM-ATT [18]	0.832	-	-	-	0.836	0.832	0.829
	DistilBert [19]	-	**0.875**	0.636	0.737	0.647	0.88	0.746
	CMC	**0.894**	0.806	**0.862**	**0.833**	**0.944**	**0.92**	**0.932**
Gossip	RoBERTa-MWSS [16]	0.80	-	-	-	0.80	-	-
	SAFE[5]	0.838	-	-	-	0.857	0.937	0.895
	Spotfake+[4]	0.856	-	-	-	-	-	-
	TM [17]	0.842	-	-	-	0.896	-	-
	LSTM-ATT [18]	0.842	-	-	-	0.839	0.842	0.821
	DistilBert [19]	-	0.805	0.527	0.637	0.866	0.960	0.911
	CMC	**0.893**	**0.826**	**0.657**	**0.692**	**0.920**	**0.963**	**0.935**

Table 3. Ablation studies on Weibo. 'S' means single-modal, while 'M' represents multi-modal.

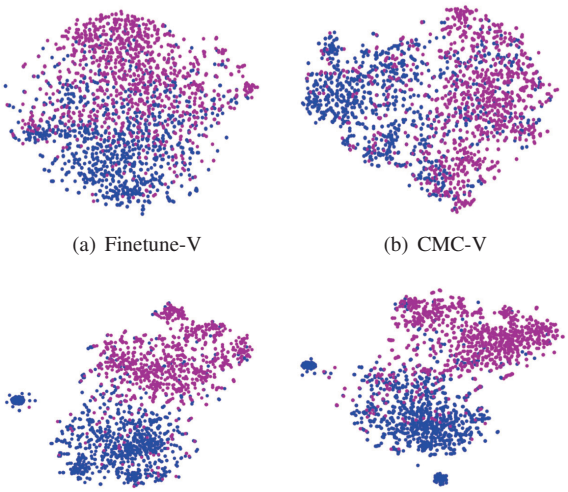
Method	Modal	Acc	Fake News			Real News		
			Prec	Rec	F1	Prec	Rec	F1
Finetune-V	S	0.594	0.590	0.617	0.603	0.597	0.570	0.583
Finetune-T	S	0.898	0.905	0.867	0.898	0.870	0.906	0.899
CMC-V	S	0.689	0.666	0.764	0.711	0.722	0.614	0.664
CMC-T	S	0.904	0.936	0.869	0.898	0.874	0.941	0.900
CMC-shared	M	0.896	0.911	0.88	0.895	0.876	0.914	0.898
CMC	M	0.908	0.940	0.891	0.900	0.883	0.945	0.907

CMC-shared: The variant of CMC that applies a shared representation between two single-modal networks. Specifically, the features extracted by two single-modal networks are integrated by the fusion mechanism, and the whole framework is trained in one stage. We use the same fusion mechanism (BLOCK [7]) as CMC for a fair comparison.

We can learn from Table 3 that the performances of both two single-modal networks are improved after applying the cross-modal distillation loss. Specifically, CMC-V surpasses Finetune-V by 9.5% accuracy while CMC-T improves Finetune-T with 0.6% accuracy. It's evident that the textual network is more powerful and can improve the performance of the visual network by a cross-modal distillation loss.

Moreover, we present the t-SNE [20] visualizations of features that are learned by Finetune-V, Finetune-T, CMC-V, and CMC-T on the test dataset of Weibo in Fig. 2 . The dots with the same color mean that they are within the same label. From Fig. 2 we can see that in both CMC-V and CMC-T, the same-label dots are comparatively closer than the different-label dots. This phenomenon reveals that the extracted features in CMC-V and CMC-T are more discriminative.

By comparing CMC-V, CMC-T, and CMC in Table 3, we observe that CMC consistently outperforms two single-modal networks CMC-V and CMC-T. It demonstrates that the fusion mechanism helps CMC to preserve more plentiful information than the single-modal methods.



(a) Finetune-V (b) CMC-V (c) Finetune-T (d) CMC-T

Fig. 2. t-SNE visualization of learned features on the test dataset of Weibo. Finetune-V and Finetune-T represent single-modal networks trained with a cross-entropy loss to perform fake news detection tasks separately. CMC-V and CMC-T denote the same two single-modal networks but are trained mutually with both the cross-entropy loss and the proposed cross-modal distillation loss.

Furthermore, we examine the performance of CMC-shared and CMC, which are both multi-modal methods. We observe from Table 3 that CMC outperforms CMC-shared consistently. We also found that CMC-shared shows little advantage over Finetune-T and CMC-T, which reveals that the multi-modal knowledge in CMC-shared is not well-utilized. This phenomenon demonstrates the effectiveness of our proposed method over counterparts with shared representations.

4. CONCLUSION

In this paper, we propose a two-stage multi-modal fake news detection framework called CMC to collaboratively train two single-modal networks and transfers the cross-modal feature correlations by a novel distillation method. In the mutual training stage of CMC, the textual and visual networks perform fake news detection tasks mutually with a closely intervened training process by each other. The cross-modal distillation loss is introduced to improve the capacity of single-modal networks by the feature correlations from the other peer. In the fusion training stage of CMC, the parameters of two single-modal networks are fixed, and a fusion mechanism is trained to further improve the performance by utilizing the discriminative information from different modalities. Our experimental results on three widely-used fake news databases show that CMC significantly outperforms the performance of existing state-of-the-art methods.

5. REFERENCES

[1] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 518–527.

[2] SM Shifath, Mohammad Faiyaz Khan, Md Islam, et al., "A transformer based approach for fighting covid-19 fake news," *arXiv preprint arXiv:2101.12027*, 2021.

[3] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference*, 2019, pp. 2915–2921.

[4] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 13915–13916.

[5] Xinyi Zhou, Jindi Wu, and Reza Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 354–367.

[6] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[7] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8102–8109.

[8] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101*, 2019.

[9] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.

[10] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.

[11] 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*, 2017, pp. 795–816.

[12] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, 2018.

[13] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 39–47.

[14] 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*, 2018, pp. 849–857.

[15] 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*, vol. 58, no. 1, pp. 102437, 2021.

[16] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu, "Leveraging multi-source weak social supervision for early detection of fake news," *arXiv preprint arXiv:2004.01732*, 2020.

[17] Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao, "Explainable tsetlin machine framework for fake news detection with credibility score assessment," *arXiv preprint arXiv:2105.09114*, 2021.

[18] Jun Lin, Glenna Tremblay-Taylor, Guanyi Mou, Di You, and Kyumin Lee, "Detecting fake news articles," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3021–3025.

[19] Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta, "Like article, like audience: Enforcing multimodal correlations for disinformation detection," *arXiv preprint arXiv:2108.13892*, 2021.

[20] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.