Bootstrapping Multi-View Representations for Fake News Detection

Qichao Ying[1], Xiaoxiao Hu[1], Yangming Zhou[1], Zhenxing Qian[1]*, Dan Zeng[2], Shiming Ge[3]

[1] Fudan University
[2] Shanghai University
[3] Chinese Academy of Sciences
{qcying20@,xxhu21@m.,ymzhou21@m.,zxqian@}fudan.edu.cn, dzeng@shu.edu.cn, geshiming@iie.ac.cn

Abstract

Previous researches on multimedia fake news detection include a series of complex feature extraction and fusion networks to gather useful information from the news. However, how cross-modal consistency relates to the fidelity of news and how features from different modalities affect the decision-making are still open questions. This paper presents a novel scheme of Bootstrapping Multi-view Representations (BMR) for fake news detection. Given a multi-modal news, we extract representations respectively from the views of the text, the image pattern and the image semantics. Improved Multi-gate Mixture-of-Expert networks (iMMoE) are proposed for feature refinement and fusion. Representations from each view are separately used to coarsely predict the fidelity of the whole news, and the multimodal representations are able to predict the cross-modal consistency. With the prediction scores, we reweigh each view of the representations and bootstrap them for fake news detection. Extensive experiments conducted on typical fake news detection datasets prove that BMR outperforms state-of-the-art schemes.

Introduction

Fake news are specified as the news that are intentionally fabricated and can be verified as false. In many cases, it is difficult for people to identify fake news. Manually removing false news one by one is laborious and expensive. Automatic Fake News Detection (FND) has become a hot research topic (Allein, Moens, and Perrotta 2021; Shu et al. 2020a; Xue et al. 2021). The FND techniques can effectively help analyze the probability of misconducting information.

The general paradigm of machine-learning-based fake news detection is to transform the news into a multidimensional latent representation, and identify the fidelity of the news using binary classification. Existing methods can be classified into three categories, namely, unimodal fake news detection (Bhattarai, Granmo, and Jiao 2021; Shu et al. 2020b), multimodal fake news detection (Khattar et al. 2019; Xue et al. 2021) and dynamic fake news detection, i.e, utilizing propagation graph (Qian et al. 2018; Xu et al. 2022) or knowledge graph (Abdelnabi, Hasan, and Fritz 2022; Sun et al. 2022). However, most posts in the online social networks are in the multimodal style. Hence, the detection based on unimodal features is far from enough, and accordingly, we focus on multimodal fake news detection.

Many existing multimodal FND schemes use the textual and visual features as integrated representations (Khattar et al. 2019; Singhal et al. 2020; Chen et al. 2019; Allein, Moens, and Perrotta 2021). Nevertheless, the disentanglement of features from different views has not been thoroughly investigated. In many cases, the models are at a black-box level, in which the network designs cannot explicitly highlight the most contributive components (Wang et al. 2018; Singhal et al. 2020). Besides, many recent works solely rely on cross-modal correlation, i.e., how semantic meaning of the image aligns with the text, to generate fused features (Wei et al. 2022; Chen et al. 2022), but we argue that cross-modal correlation not necessarily play a critical role. Fig. 1 provides four examples respectively from the well-known English GossipCop (Singhal et al. 2020) and the Chinese Weibo (Jin et al. 2017) dataset, where both fake and real news more or less contain cross-modal correlations. Therefore, clarifying the roles of both unimodal and cross-modal features is vital for improving FND.

Aiming at addressing these issues, we propose an effective scheme that Bootstraps Multi-view Representations (BMR) for fake news detection. Given a multi-modal news, we extract representations respectively from the views of the text, the image pattern and the image semantics. Improved Multi-gate Mixture-of-Expert networks (MMoE) (Ma et al. 2018), denoted as iMMoE, are proposed for feature refinement and fusion. Representations from each view are separately used to coarsely predict the fidelity of the whole news, where the prediction scores are used for adaptive feature reweighting. Cross-modal consistency learning further implicitly guides multimodal representation refinement, and we explicitly disentangle correlation from other cross-modal information by introducing an independent representation. Finally, we bootstrap multi-view features for refined fake news detection. In the examples shown in Fig. 1, BMR not only predicts the fidelity of the news, but also gives confidence scores based on each view, which provides a new way of understanding how different roles act in multimodal FND.

The contributions of this paper are three-folded:

- We propose a novel fake news detection scheme of generating multi-view representations, understanding their in-

*Corresponding author: Zhenxing Qian.

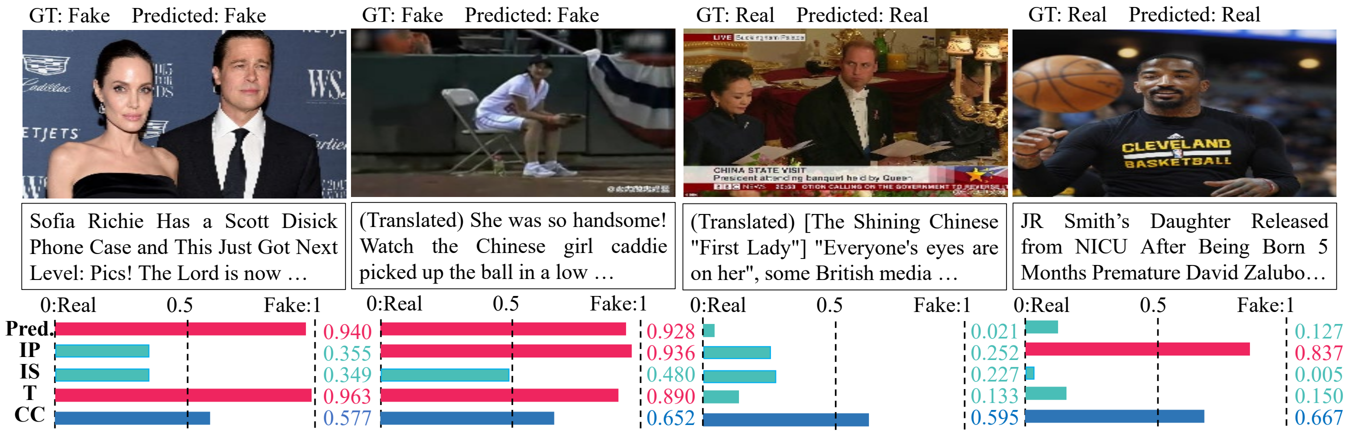


Figure 1: Examples of fake news detection result of BMR on Weibo (left two) and GossipCop (right two). The Chinese sentences from Weibo are translated here for references. Row "IP" (Image Pattern), "IS" (Image Semantics), "T" (Text) respectively show the single-view prediction results provided by BMR, which suggests the dubious parts of the news. Row "CC" (Cross-modal Consistency) shows the predictions on cross-modal consistency and Row "Pred." provides the ultimate FND results.

dividual importance, and optimizing the fused features.

- We propose to disentangle information within unimodal and multimodal features by single-view prediction and cross-modal consistency learning, which are then adaptively reweighed and bootstrapped for better detection.
- The proposed BMR detection not only outperforms state-of-the-art multimodal FND schemes on popular datasets, but also provides a mechanism for interpreting the contributions of different representations.

Related Works

Unimodal Fake News Detection. Both the language-based and the vision-based fake news detections have received extensive attention. TM (Bhattarai, Granmo, and Jiao 2021) utilizes lexical and semantic properties of the text to detect fake news. MWSS (Shu et al. 2020b) exploits multiple weak signals from different sources from user and content engagements. Jin et al. (Jin et al. 2016) find that there are noticeable differences in the distribution of images between real news and fake news. Cao et al. (Cao et al. 2020) suggest that typical methods for image manipulation detection (Chen et al. 2021) are useful in unveiling traces for news tampering. Besides, Qi et al. (Qi et al. 2019) jointly use the spatial domain and frequency domain features of the news for forensics. However, for the multimodal news, these approaches are unable to detect cross-modal correlations.

Multimodal Fake News Detection. The critical issue of multimodal fake news detection is to align linguistic and vision representations. SpotFake (Singhal et al. 2020) integrates pretrained XLNet and ResNet for feature extraction. SAFE (Zhou, Wu, and Zafarani 2020) feeds the relevance between news textual and visual information into a classifier to detect fake news. EANN (Wang et al. 2018) introduces an additional discriminator to classify news events so as to suppress the impact of specific events on classification. MCNN (Xue et al. 2021) also incorporates textual semantic features, visual tampering features, and similarity of textual and visual information, but the aggregation process only concatenates all features. Hence, how each of these multi-view features affects the predictions cannot be measured. In comparison, we explicitly reweigh multi-view representations based on single-view classifications and adaptively bootstrap them for fake news detection.

Besides, some methods propose to use cross-modal correlation learning for fake news detection. CAFE (Chen et al. 2022) uses a VAE to compress images and texts representations and measures cross-modal consistency based on their Kullback-Leibler (KL) divergence. The consistency score then linearly adjusts the weight of unimodal and multimodal features before final classification. Likewise, CMC (Wei et al. 2022) includes a two-staged network that train two uni-modal networks to learn cross-modal correlation by contrastive learning, and then finetune the network for fake news detection. However, CAFE severely penalizes unimodal features when the consistency is high, which might be doubtful in some news. CMC does not adaptively suppress or augment multi-view representations, and the fine-tuning stage may erase the cross-modal knowledge learned in the first stage. Therefore, we propose an improved mechanism to better utilize cross-modal consistency learning for FND.

Proposed Approach

Fig. 2 depicts the pipeline of BMR, which contains four stages, i.e., Multi-view Feature Extraction, Refine & Fusion, Disentangling & reweighting, and Bootstrapping. In the first three stages, there are four branches corresponding to the representations from four views, including the Image Pattern Branch, the Image Semantics Branch, the Text Branch, and the Fusion Branch.

Multi-View Feature Extraction

Let the input multimodal news be $\mathcal{N} = [\mathbf{I}, \mathbf{T}] \in \mathcal{D}$, where $\mathbf{I}, \mathbf{T}, \mathcal{D}$ are the image, the text and the dataset, respectively. We begin with extracting coarse multi-view representations,

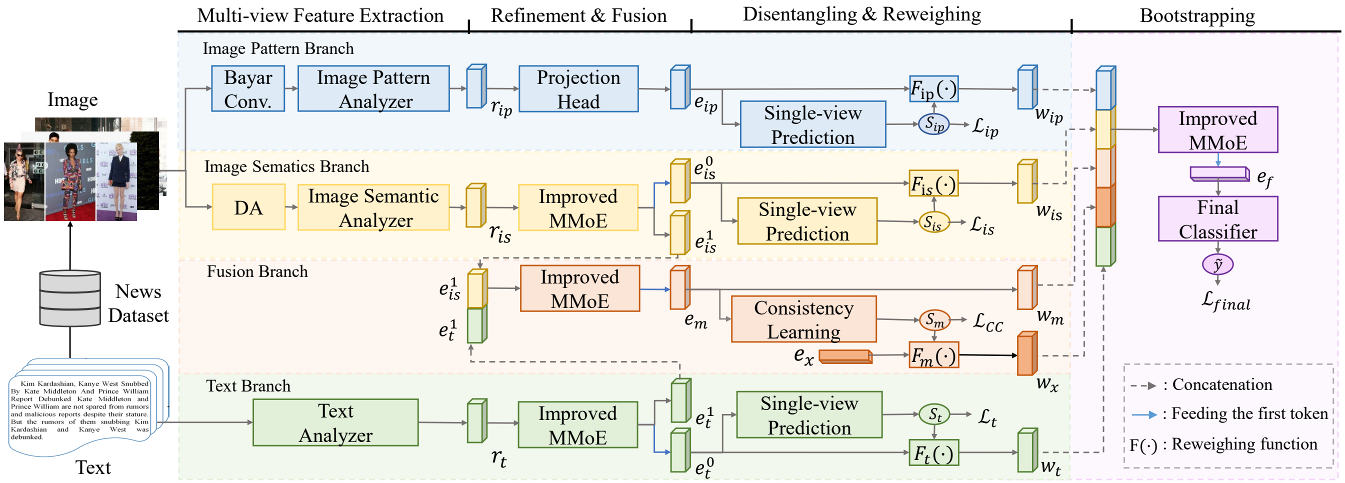


Figure 2: The network architecture of BMR. Representations are respectively extracted from the views of text, image pattern and image semantics. Improved MMoE, together with single-view predictions and cross-modal consistency learning, jointly guides unimodal feature refinement and cross-modal feature generation. The final decision is made upon bootstrapping these multi-view representations.

including the image pattern, the image semantics, and the text. We denote these representations as r_{ip}, r_{is} and r_t. Our view is that the general distribution of the image and the tiny traces left by tampering or compression can be helpful to expose fake news. Therefore, we explicitly separate the feature learning paradigm from image pattern and semantics. In the Image Pattern Branch, we use InceptionNet-V3 (Szegedy et al. 2016) as the image pattern analyzer. In the Image Semantic Branch, we use Masked Autoencoder (MAE) as the image semantics analyzer to extract r_{is}. In the Text Branch, we use BERT (Devlin et al. 2018) to extract r_t.

We explicitly disentangle image pattern and semantics learning from two aspects. 1) Architectural difference. While CNNs are famous for pattern recognition, the transformer models (He et al. 2022; Gabbay, Cohen, and Hoshen 2021) based on masked language/image modeling are reported to be great in establishing long-range attention. 2) Input difference. We include a BayarConv (Bayar and Stamm 2018) in the IP branch and Data Augmentation (DA) methods, e.g., flipping, color adjustment, in the IS branch. BayarConv augments details and suppresses the main component of **I**, while DA encourages the networks to learn more robust semantic features that are ignorant to trivial changes.

Feature Refinement and Fusion with iMMoE

In the Refine & Fusion Stage, we refine the representations extracted from the first stage. In the Image Pattern Branch, the representation r_{ip} is projected to a new representation e_{ip}, where the classification head of the InceptionNet-V3 is replaced with an MLP-based projection head. The output size of e_{ip} equals to that of a single token representation from MAE. In the Image Semantics Branch and the Text Branch, we propose the improved MMoE (iMMoE) network to refine r_{is} and r_t. We also generate a new representation e_m in the Fusion Branch.

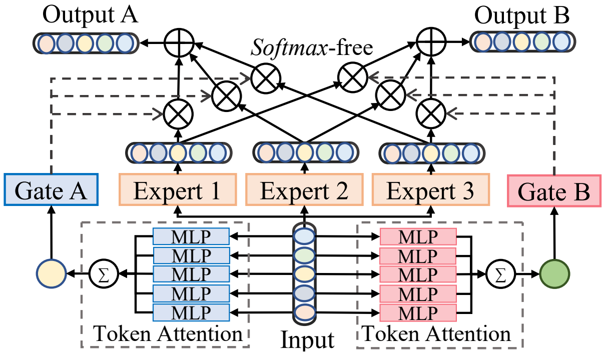


Figure 3: Network architecture of improved MMoE.

As sketched in Fig. 3, the proposed iMMoE contains several experts, gates and token attentions that produces features in separate branches. As summarized in Eq. (1), the MMoE network (Ma et al. 2018) is originally designed to model multi-task relationships from data by sharing the expert across all tasks. The input x is equally sent into n expert networks, and k gates adaptively weigh the outputs of the experts as the final output using *softmax* function. n is a hyper-parameter and k is the amount of down-stream tasks.

$$x^k = \sum_{i=1}^{n} \textit{softmax}(G_i^k(x)) \cdot \; E_i(x), \tag{1}$$

where E_i and G_i^k are the i_{th} expert and the i_{th} output of the gate for task k. In BMR, we treat the mining of multi-view unimodal representations and cross-modal feature fusion as different subtasks. They should share common features and reserve their own distinctive features. We improve MMoE network in two aspects. First, we use token attention to compute the importance scores of each token representation using a *weight-shared* MLP, and perform dimension reduction by aggregating all token representations into one according to the scores. The aggregated representation is then sent into the gate to compute the weights for each expert. Second, we find that the *softmax* functions in the gates require that all experts must contribute positively to all outputs. The restriction is not necessary according to our experiments. We lift the *softmax* constraints and allow the weights to be negative or greater than one. We revise Eq. (1) as Eq. (2) for iMMoE,

$$x^k = \sum_{i=1}^{n}(G_i^k(\sum_{j=1}^{t} MLP_k(x)) \cdot \ E_i(x)), \tag{2}$$

where t is the amount of tokens and MLP_k denotes the token attention for task k. $k \in [1, 2]$.

With the iMMoE network, we refine r_{is} and r_t into the features $[e_{is}^0, e_{is}^1]$ and $[e_t^0, e_t^1]$, respectively. e_{is}^0 and e_t^0 are preserved for single-view prediction. Meanwhile, the e_{is}^1 and e_t^1 are jointly fed into the iMMoE net in the Fusion Branch to generate a multimodal feature e_m, which is preserved for cross-modal consistency learning and bootstrapping.

Disentangling & Reweighting

After the refinement and fusion, we obtain the new representations from four branches. Next, we handle representations of image pattern, image semantics and text via single-view prediction, while processing the fused representation in the Fusion Branch for consistency learning.

Single-View Prediction and Reweighting. Single-view prediction uses the first token of e_{is}^0, the first token of e_t^0 and e_{ip}, to predict the fidelity of $\mathcal{N}$. We design this module based on two considerations. On one hand, many fake news contains obvious abnormality in either image noise, distribution or text independently. Therefore, making predictions on $\mathcal{N}$ with single-view features is empirically feasible. On the other hand, the confidence of each single-view prediction can be projected into weights that adaptively reweigh the representations, serving as *modality-wise attentions*. Therefore, we use MLP to project each single-view representation into scores, and use another MLP, i.e., $F(\cdot)$ in Fig. 2, to project the scores into weights. Take the generation of S_{ip} and w_{ip} as an example,

$$\begin{aligned} S_{ip} &= MLP_{ip}(e_{ip}), \\ w_{ip} &= Sigmoid\left(F_{ip}\left(S_{ip}\right)\right) \cdot e_{ip}, \end{aligned} \tag{3}$$

where $MLP_{ip}(\cdot)$ denotes the MLP-based single-view predictor. In the same way, we generate paired prediction scores and reweighed representations $\{S_{is}, w_{is}\}$, $\{S_m, w_m\}$ and $\{S_t, w_t\}$. The reweighed representations are then ready for bootstrapping. Compared to the widely used channel-wise or spatial-wise attentions (Woo et al. 2018), disentangling & reweighting stage in BMR better explicitly reserves the most useful information in each view of the news.

Cross-Modal Consistency Learning. Inspired by several recent works (Chen et al. 2022; Wei et al. 2022), we guide the multimodal feature fusion in the Fusion Branch with cross-modal consistency learning. Specifically, we train BMR to predict whether a given text-image pair matches. We begin with crafting a new dataset $\mathcal{D}' = [\mathcal{D}_{real}, \mathcal{D}_{syn}]$ on the basis of $\mathcal{D}$, where news with correlated texts and images are labeled $y' = 1$, otherwise $y' = 0$. Fig. 4 shows three training data for cross-modal consistency learning, where the positive examples are directly the real news borrowed from $\mathcal{D}$ and the negative examples are synthesized "news" by arbitrarily combining images and texts from different real news. Though there can be cases that some news in $\mathcal{D}_{real}$ do not contain cross-modal consistency, the possibility of text-image pairs from $\mathcal{D}_{syn}$ having consistency is much lower in nature. Therefore, the model can still learn the correlation based on such a hybrid dataset. We feed BMR with $\mathcal{N}' = [\mathbf{I}', \mathbf{T}'] \in \mathcal{D}'$. After the corresponding e_m is calculated, we use an MLP-based predictor to output the consistency score S_m, and expect the score to be close to the label y'. The training process of cross-modal consistency learning only activates a part of the modules in BMR. The task can be in parallel learned with the main task.

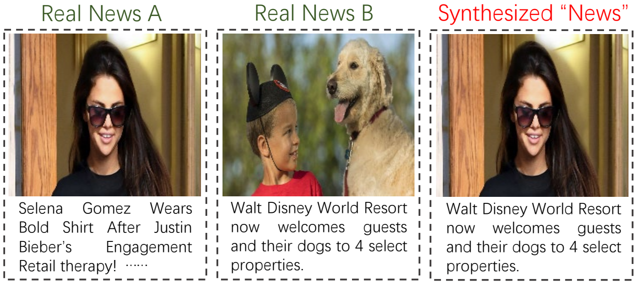

Real News A
Real News B
Synthesized “News”
Selena Gomez Wears Bold Shirt After Justin Bieber's Engagement Retail therapy! ……
Walt Disney World Resort now welcomes guests and their dogs to 4 select properties.
Walt Disney World Resort now welcomes guests and their dogs to 4 select properties.

Figure 4: Training data for cross-modal consistency learning. Real news borrowed from typical FND datasets are marked positive and synthesized news by combining image and text from different real news are marked negative.

Reweighting Multimodal Representation. The predicted consistency scores on $\mathcal{N}$ also adjust the multimodal representations. We consider that multimodal features should not merely represent cross-modal correlation. There are also many other factors, e.g., joint distribution, emotional difference, etc., that can help decide whether the news is fake. Therefore we refrain from weighing e_m using S_m. Instead, we explicitly disentangle correlation from other cross-modal information by letting S_m reweigh an independent trainable token e_x that represents *cross-modal irrelevance*. $w_x = e_x \cdot \ F_m(S_m)$ and $w_m = e_m$, which are both considered multimodal representations for bootstrapping. Compared to previous works, we do not separate the cross-modal learning with the detecting stage (Wei et al. 2022) or reweighting both unimodal and multimodal representations by correlation score (Chen et al. 2022).

Bootstrapping Stage and Loss Function

The multi-view representations $[w_{is}, w_{ip}, w_m, w_x, w_t]$ are bootstrapped using another iMMoE, which further refines the information critical for the decision-making. The final MLP-based classifier gets the first token of the output e_f to predict $\tilde{y}$, which is expected to be close to the label y.

Fake news detection is a binary classification problem. We apply the BCE loss between the ground-truth label y and

Algorithm 1: PyTorch-style training code of BMR

```
Input: Dataset: $\mathcal{D}$, Training epochs: $N$, Amount of news for constructing $\mathcal{D}'$: $k$
Output: Model parameters: $\Theta$.
Sample $k/2$ real news from $\mathcal{D}$ and store them into $\mathcal{D}'$ as positive examples.
for i in range($k$/4): do
    Sample $\mathcal{N}_1 = [\mathbf{I}_1, \mathbf{T}_1], \mathcal{N}_2 = [\mathbf{I}_2, \mathbf{T}_2]$ from $\mathcal{D}$.
    Synthesize pseudo-news by shuffling information in $\mathcal{N}_1$ and $\mathcal{N}_2$. $\mathcal{N}_3 = [\mathbf{I}_1, \mathbf{T}_2], \mathcal{N}_4 = [\mathbf{I}_2, \mathbf{T}_1]$.
    Store $\mathcal{N}_3, \mathcal{N}_4$ into $\mathcal{D}'$ as negative examples.
end for
for i in range(N): do
    Sample $(\mathbf{I}, \mathbf{T}, y)$, $(\mathbf{I}', \mathbf{T}', y')$ from $\mathcal{D}, \mathcal{D}'$
    $S_m = BMR(\mathbf{I}', \mathbf{T}', train\_consist = True)$
    Compute loss using $\mathcal{L}_{BCE}(y', S_m)$.
    $\mathbf{T}$ ="No text provided." if $len(\mathbf{T}) < 5$.
    $\mathbf{I} = zeros\_like(\mathbf{I})$ if $\mathbf{I}.rows < 64$ or $\mathbf{I}.cols < 64$
    $[\tilde{y}, S_t, S_{is}, S_{ip}] = BMR(\mathbf{I}, \mathbf{T}, train\_consist = False)$.
    Compute loss using $\mathcal{L}_{BCE}(y, \tilde{y})$, $\mathcal{L}_{BCE}(y, S_t)$, $\mathcal{L}_{BCE}(y, S_{is})$, $\mathcal{L}_{BCE}(y, S_{ip})$.
    Update parameters in $\Theta$ using Adam optimizer.
end for
```

the predicted scores $\tilde{y}$, as well as the coarse classification results S_{ip}, S_{is}, S_t. There is an extra loss for cross-modal consistency training, where we also apply the BCE loss between the ground-truth crafted label y' and S_m. Therefore, the losses are defined as $\mathcal{L}_{CC} = \mathcal{L}_{BCE}(y', S_m)$, $\mathcal{L}_{final} = \mathcal{L}_{BCE}(y, \tilde{y})$, $\mathcal{L}_T = \mathcal{L}_{BCE}(y, S_t)$, $\mathcal{L}_{ip} = \mathcal{L}_{BCE}(y, S_{ip})$ and $\mathcal{L}_{is} = \mathcal{L}_{BCE}(y, S_{is})$. The single view FND classification loss is the aggregate of $\mathcal{L}_{is}$, $\mathcal{L}_{ip}$ and $\mathcal{L}_t$, namely,

$$\mathcal{L}_{coarse} = (\mathcal{L}_{is} + \mathcal{L}_{ip} + \mathcal{L}_t)/3, \tag{4}$$

The total loss for BMR is defined as follows.

$$\mathcal{L} = \mathcal{L}_{final} + \alpha \cdot \mathcal{L}_{coarse} + \beta \cdot \mathcal{L}_{CC}, \tag{5}$$

where α and β are hyper-parameters. We provide the pseudo-code of BMR in Algo. 1.

Implementation Details

We use the "mae-pretrain-vit-base" model for image processing, the "bert-base-chinese" model for Chinese dataset, and the "bert-base-uncased" model for English dataset. The hidden sizes of both MAE and BERT are 768. BERT and MAE are kept frozen. All MLPs in BMR contain one hidden layer, a BatchNorm1D and an ELU activation. We use the single-layer transformer blocks defined in ViT (Gabbay, Cohen, and Hoshen 2021) to implement the experts in iMMoE. Each iMMoE network holds three experts. We use Adam optimizer with the default parameters. The batch size is 24 and the learning rate is 1×10^{-4} with cosine annealing decay. Images are resized into size 224×224, and the max length of text is set as 197. BMR is also designed to be compatible with unimodal news by feeding a *zero matrix* as the image or *"No text provided"* as the text. Similar to EANN (Wang et al. 2018), we increase the quality of the datasets by replacing images smaller than 64×64 and texts less than five words with the above-mentioned placeholders.

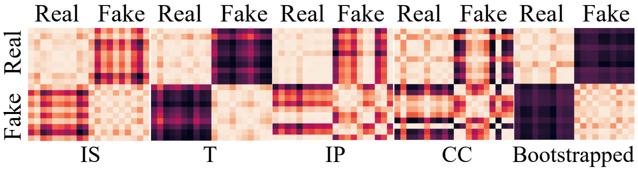


Figure 5: Heatmap visualization. Each cell in the heat maps represents the paired cosine similarity between the 64-dim representations respectively from the final classifier and the single-view predictors. Tests are done on Weibo.

Experiments

Experimental Setups

We use Weibo (Jin et al. 2017), GossipCop (Shu et al. 2020a) and Weibo-21 (Nan et al. 2021) for training and testing. Weibo contains 3749 real news and 3783 fake news for training, 1000 fake news and 996 real news for testing. GossipCop contains 7974 real news and 2036 fake news for training, 2285 real news and 545 fake news for testing. Weibo-21 is a newly-released dataset that contains 4640 real news and 4487 fake news in total, and we split it into training and testing data at a ratio of $9 : 1$. Though MediaEval (Boididou et al. 2018) and Politifact (Singhal et al. 2020) are also popular datasets, they contain only 460 images and 381 posts in the training set, and neural nets can easily overfit with such small amount of data. We train BMR on each dataset five times with different initial weights and report the averaged best performance. The hyper-parameters α and β are empirically set as $\alpha = 1$, and $\beta = 4$. Experiments show that the BCE loss term converges much quicker than consistency term, which lets BMR neglect the consistency. We find that $\beta = 4$ best mitigates this as an acceleration.

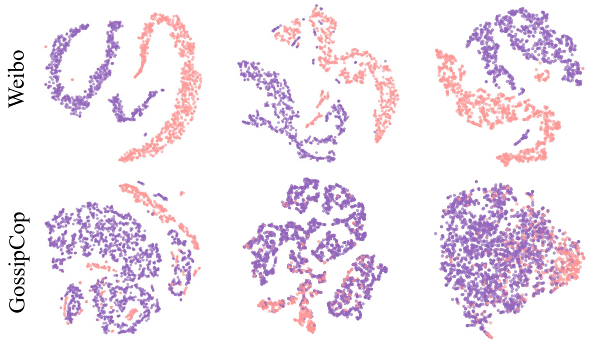
According to the theory in (He and Garcia 2009), we determine a fixed threshold for each dataset based on its distribution. Since the ratio of real news versus fake news in GossipCop training set is close to 4:1, we set the threshold for GossipCop as 0.80. Similarly, the thresholds for Weibo and Weibo-21 are set as 0.50. For real-world applications, we can set a default threshold as 0.50.

Performance Analysis

Heatmap Visualization. In Fig. 5, we measure the discriminative capability of the multi-view representations using heatmap visualization. We arbitrarily select ten real news and ten fake news, and then compute paired similarity between the 64-dim representations from the final classifier and the single-view predictors. From the figures, the bootstrapped representation shows strong discriminative capability where intra-class similarity and inter-class difference are both noticeable. Some other views of representations also have decent borders, but the heatmap for cross-modal consistency tends to be much messier, showing that it is not practical to directly use the consistency for FND.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	EANN* (Wang et al. 2018)	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	MCNN (Xue et al. 2021)	0.846	0.809	0.857	0.832	0.879	0.837	0.858
	MCAN (Wu et al. 2021)	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	CAFE* (Chen et al. 2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	CMC (Wei et al. 2022)	0.893	**0.940**	0.869	0.899	0.876	**0.945**	**0.907**
	BMR (Proposed)	**0.918**	0.882	**0.948**	**0.914**	**0.942**	0.870	0.904
GossipCop	EANN* (Wang et al. 2018)	0.864	0.702	0.518	0.594	0.887	0.956	0.920
	Spotfake* (Singhal et al. 2020)	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	DistilBert (Allein, Moens, and Perrotta 2021)	0.857	0.805	0.527	0.637	0.866	0.960	0.911
	CAFE* (Chen et al. 2022)	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	CMC (Wei et al. 2022)	0.893	**0.826**	**0.657**	**0.692**	**0.920**	0.963	0.935
	BMR (Proposed)	**0.895**	0.752	0.639	0.691	**0.920**	**0.965**	**0.936**
Weibo-21	EANN* (Wang et al. 2018)	0.870	0.902	0.825	0.862	0.841	**0.912**	0.875
	SpotFake* (Singhal et al. 2020)	0.851	**0.953**	0.733	0.828	0.786	0.964	0.866
	CAFE* (Chen et al. 2022)	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	BMR (Proposed)	**0.929**	0.908	**0.947**	**0.927**	**0.946**	0.906	**0.925**

Table 1: Comparison between BMR and state-of-the-art multimodal fake news detection schemes on Weibo, GossipCop and Weibo-21. *: open-sources. The best performance is highlighted in bold red and the follow-up is highlighted in bold blue.



(a) BMR (b) EANN (c) CAFE

Figure 6: TSNE visualization of mined features on the test set. Dots with the same color are within the same label.

	BMR	EANN	SpotFake	CAFE
Params	94.39M	143.70M	124.37M	0.68M
FLOPS	18.42G	19.63G	30.42G	0.01G

Table 2: Comparison of trainable parameters and computational speed. FLOPs: amount of floating point arithmetics.

Comparisons. Table 1 shows the average precision, recall, and accuracy of BMR on Weibo and GossipCop. We use Accuracy, Precision, Recall, and F1 score for comprehensive performance measurements. The results are promising, with 91.8% average accuracy on Weibo, 92.9% on Weibo-21 and 89.5% on GossipCop. We further compare BMR with state-of-the-art methods. BMR outperforms the other methods on the datasets. Besides, we rank either $1st$ or $2nd$ on Recall and F1 score of fake news on all of the datasets. CMC has a very close performance with BMR on GossipCop. However, on Weibo, BMR outperforms CMC by a more noticeable 2.5%. We have also trained three open-sourced methods on Weibo-21 and compared them with BMR. The result shows that BMR leads by 4.2% in overall accuracy and 3.2% in Recall of fake news, which proves the effectiveness of BMR.

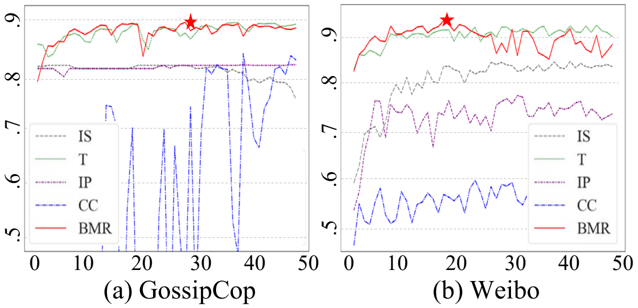
Moreover, we present the t-SNE visualizations of features in Fig. 6, which are learned by BMR, SpotFake and CAFE on the test set of GossipCop. In BMR, the dots representing fake news are comparatively farther from differently-labeled dots, and there are fewer outliers. This indicates that the extracted features in BMR are more discriminative.

Computational Complexity. It costs around 12 hours to train BMR for 50 epochs on a single NVIDIA RTX 3090 GPU. The computational complexity study in Table 2 shows that BMR requires less trainable parameters compared to EANN and SpotFake. CAFE requires amazingly low computational complexity by using simple auto-encoders and MLPs, but the performance is not good enough.

Ablation Studies

Contribution of Each View. Fig. 7 and Table. 3 show the curves for testing accuracy and ultimate training loss. We study the contribution of each view on the datasets. First, we find that $\mathcal{L}_{ip}$, $\mathcal{L}_{is}$ and $\mathcal{L}_{CC}$ do not fully converge on GossipCop. Therefore we closer scrutinized GossipCop dataset and surprisingly find that close to 50% images are celebrity faces that are even hard to distinguish for many human viewers. However, BMR still defeats all compared FND schemes in the overall accuracy, which indicates that FND on GossipCop relies severely on text and images might contribute marginally. Second, we study whether there is under-fitting, i.e., single-view predictions struggling with the ground truth y, or over-fitting, i.e., networks remembering the training set. In Fig. 7, the testing accuracies remain high and the curves go into plateau rather than drop even though we train 50 epochs. It indicates mild temptation to overfitting.

Removal of Some Views. In the upper part of Table 4, we feed BMR with different combinations of views of representations. For example, in row three, we only feed BMR with the image, preserve the Image Pattern Branch and the Image Semantics Branch, while disabling the rest of the modules.



(a) GossipCop (b) Weibo

Figure 7: Curves of testing accuracy (y-axis) versus training epochs (x-axis). “CC” curve represents using cross-modal consistency as the judgement of fake news detection.

Dataset	$\mathcal{L}_{final}$	$\mathcal{L}_{CC}$	$\mathcal{L}_{t}$	$\mathcal{L}_{is}$	$\mathcal{L}_{ip}$
Weibo	0.005	0.084	0.007	0.026	0.133
GossipCop	0.005	0.540	0.006	0.511	0.493
Weibo21	0.002	0.124	0.003	0.067	0.012

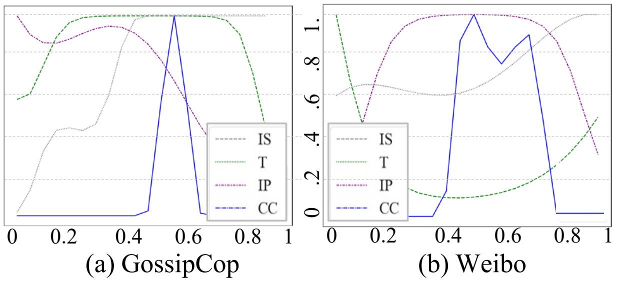
Table 3: Averaged ultimate training loss of each term.

Test	Accuracy	F1 Score	
		Fake News	Real News
IP	0.789	0.774	0.806
IS	0.838	0.860	0.809
IS+IP	0.857	0.874	0.833
T	0.870	0.872	0.868
IP+T	0.878	0.876	0.876
IS+T	0.881	0.888	0.873
IS+IP+T	0.886	0.888	0.884
S_m reweigh. Multi-view.	0.864	0.854	0.809
w/o Feature Reweigh.	0.887	0.880	0.892
w/o Coarse Class.	0.895	0.903	0.887
using ViT Blocks for Refine.	0.905	0.910	0.900
w/o Cross. Correlat.	0.905	0.906	0.902
using ViT instead of MAE	0.910	0.911	0.900
w/o improving MMoE	0.907	0.910	0.899
BMR	**0.918**	**0.914**	**0.904**

Table 4: Ablation study on bootstrapping multi-view representations and network design. The tests are done on Weibo.

We find that if we bootstrap three multi-view unimodal representations without introducing the fusion branch, the averaged accuracy is 88.6%, which is 3.2% worse than BMR. Though we find that the contribution of text is large for FND, merely using texts only guarantee 87.0% accuracy.

Cross-Modal Consistency Learning. In Table 4, we mimic CAFE that reweighs e_{ip}, e_{is}, e_t using S_m. We surprisingly find that the result is even worse than not using multimodal features. It suggests that granting priority to cross-modal consistency rather than unimodal representations is not beneficial. Second, we remove cross-modal consistency learning and therefore add no constraint on multimodal feature generation. The performance drops 1.3% accordingly. From Table 3 and Fig. 7, we also find that though the training losses for consistency learning on Weibo and Weibo-21 are both low, we fail to conduct FND simply based on cross-modal consistency, where the accuracy is less than 0.6. Since we have ruled out over-fitting, we conclude that the proposed training methodology for cross-modal consistency is feasible but the consistency can so far only play an assistant rather than a critical role in multimodal FND.



(a) GossipCop (b) Weibo

Figure 8: Curves of the learned reweighting functions $F(\cdot)$. x-axis: score, e.g., S_{ip}, y-axis: corresponding weight.

Single-View Prediction. In Table 4, we test the case of adding no constraint to the generation of multi-view representation, which results in a performance that drops 2.3%. Therefore, single-view prediction is a useful manual bias compared to blindly employing multi-branches for feature extraction. Our quantitative results show that 127 of 200 arbitrarily selected real news are correctly predicted even though some of the single-view classifiers predict fake. It suggests that the bootstrapped prediction is more tolerant to comparatively minor perturbations. Besides, if we do not explicitly reweigh the representations according to the subtasks, the accuracy will decrease by 2.9%, indicating that there would be misleading information if we do not reweigh the representations. Fig. 8 shows the learned reweighting functions $F(\cdot)$ on GossipCop and Weibo. Many of these learned functions exhibit inverted “V“ shapes. It suggests that BMR learns to penalize over-confidence for more robust bootstrapped prediction. Besides, e_x is only activated when the cross-modal consistency score is close to 0.5.

Network Design. Since we are the first to use MAE in FND, we verify the case of replacing MAE with ViT. The overall accuracy of BMR drops 0.8% results in Table. 4, which indicates that MAE is better for the task. Besides, When using iMMoE instead of separate ViT blocks, the result will increase by 1.3%, suggesting that sharing some representations is beneficial. We also find that the internal improvement made in MMoE in BMR helps gaining 1.1%.

Conclusions

We propose BMR that generates multi-view representations, understands the individual importance, and optimizes the fused features. Single-view prediction and cross-modal consistency learning are proposed to disentangle information within unimodal and multimodal features, which are then adaptively reweighed and bootstrapped for better detection results. Experiments show that BMR outperforms state-of-the-art multimodal FND schemes on popular datasets.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant U20B2051.

References

Abdelnabi, S.; Hasan, R.; and Fritz, M. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14940–14949.

Allein, L.; Moens, M.-F.; and Perrotta, D. 2021. Like Article, Like Audience: Enforcing Multimodal Correlations for Disinformation Detection. *arXiv preprint arXiv:2108.13892*.

Bayar, B.; and Stamm, M. C. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11): 2691–2706.

Bhattarai, B.; Granmo, O.-C.; and Jiao, L. 2021. Explainable Tsetlin Machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*.

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; and Li, J. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, 141–161.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14185–14193.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Chen, Y.; Sui, J.; Hu, L.; and Gong, W. 2019. Attention-residual network with CNN for rumor detection. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1121–1130.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gabbay, A.; Cohen, N.; and Hoshen, Y. 2021. An image is worth more than a thousand words: Towards disentanglement in the wild. *Advances in Neural Information Processing Systems*, 34.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9): 1263–1284.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3): 598–608.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1930–1939.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 518–527. IEEE.

Qian, F.; Gong, C.; Sharma, K.; and Liu, Y. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *IJCAI*, volume 18, 3834–3840.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020a. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.

Shu, K.; Zheng, G.; Li, Y.; Mukherjee, S.; Awadallah, A. H.; Ruston, S.; and Liu, H. 2020b. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*.

Singhal, S.; Kabra, A.; Sharma, M.; Shah, R. R.; Chakraborty, T.; and Kumaraguru, P. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13915–13916.

Sun, M.; Zhang, X.; Zheng, J.; and Ma, G. 2022. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4611–4619.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wei, Z.; Pan, H.; Qiao, L.; Niu, X.; Dong, P.; and Li, D. 2022. Cross-Modal Knowledge Distillation in Multi-Modal Fake News Detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4733–4737. IEEE.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569.

Xu, W.; Wu, J.; Liu, Q.; Wu, S.; and Wang, L. 2022. Mining Fine-grained Semantics via Graph Neural Networks for Evidence-based Fake News Detection. *arXiv preprint arXiv:2201.06885*.

Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; and Wei, L. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5): 102610.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 354–367. Springer.