Pattern Recognition Letters 195 (2025) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

CAMFND: Cross-modal adaptive-aware learning for multimodal fake news detection

Ying Guo [a,b], Yuan Li [a], Kexin Zhen [a], Bingxin Li [a], Jie Liu [a,*]

[a] North China University of Technology, Beijing, 100144, China
[b] Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Editor: Ricardo Torres

Keywords:
Fake news detection
Multimodal learning
Adaptive learning
Social media

ABSTRACT

Recently, there has been a growing focus on the automatic identification of multimodal fake news detection. A fundamental challenge of multimodal fake news detection lies in the inherent semantic ambiguity across different content modalities. Decisions stemming from distinct unimodal sources may exhibit discrepancies, potentially creating inconsistency with the collective insights derived from multimodal data fusion. To address this issue, we propose CAMFND: a cross-modal adaptive-aware learning framework for multi-modal fake news detection, aiming to reduce semantic ambiguities among different modalities. CAMFND consists of (1) a cross-modal alignment module to transform the heterogeneous unimodality features into a shared semantic space, (2) a cross-modal adaptive-interactive module to capture the semantic correlation and consistency, computed by the multi-modal gated fusion unit, (3) a cross-modal adaptive-selective module to decide the semantic meaning or bias, guided by the multi-modal semantic matching score. CAMFND enhances the fake news detection by intelligently and dynamically combining features from uni-modality and identifying correlations across different modalities. It leverages unimodal features in scenarios with low cross-modal ambiguity, while utilizing cross-modal correlations in cases of high cross-modal uncertainty. The experimental results show that CAMFND significantly surpasses prior methodologies and sets new benchmarks on both English Twitter and Chinese Weibo datasets, marking a notable advancement in performance.

1. Introduction

Research [1] indicates that more than three billion individuals rely on Weibo and Twitter as their main sources of daily information, thus social media has emerged as the dominant platform for people to share news. Despite the convenience, the absence of systematic measures to verify the credibility of news has resulted in the rapid and widespread dissemination of fake news. Consequently, fake news detection has garnered growing research interest in recent years.

Current techniques for detecting fake news can be broadly categorized into two groups: single-modal based [2,3] and multi-modal based [4,5]. To date, diverse multi-modal approaches have concentrated on distinct modality fusion techniques to detect fake news. For instance, Wang et al. [6] utilized a straightforward splicing approach to merge features from various modalities, whereas others had enhanced fusion capabilities by employing strategies such as Attention [7] and Bilinear Pooling [8]. Nowadays, semantics are proving to provide evident clues for identifying fake news. In the incorporation of semantic features, the SAFE framework [9] was the first to introduce the semantic information, while the MCNN framework [10] introduced image tampering to propose a refined model based on it.

Recently, text–image consistency [10] has proved to be a crucial indicator in detecting fake news but it is generally limited to the main modalities, and the cross-modal semantic gap is still in its fancy, as depicted in Fig. 1. Even though some existing method [8] overemphasize cross-modal fusion interactions, they are difficult to adaptively fuse and select feature vectors in high-dimensional space, which ignores global semantics and affects the utilization of correlation. Besides, the current correlation [9] between texts and images is mainly calculated using the cosine similarity-based methods. Cosine similarity performs well in simple similarity metrics, especially on high-dimensional sparse features. However, as the complexity of the cross-modal increases, it tends to ignore the semantic relationship which only focuses on the direction in the embedding space, so it does not consider the weight relationship between cross-model features. Additionally, most methods [11,12] overly rely on multimodal features and ignore the

* Corresponding author.
E-mail addresses: guoying@ncut.edu.cn (Y. Guo), dayuanzi@mail.ncut.edu.cn (Y. Li), zkx0922@mail.ncut.edu.cn (K. Zhen), libingxin@mail.ncut.edu.cn (B. Li), liujie@ncut.edu.cn (J. Liu).

https://doi.org/10.1016/j.patrec.2025.02.035
Received 9 July 2024; Received in revised form 17 January 2025; Accepted 26 February 2025
Available online 12 May 2025

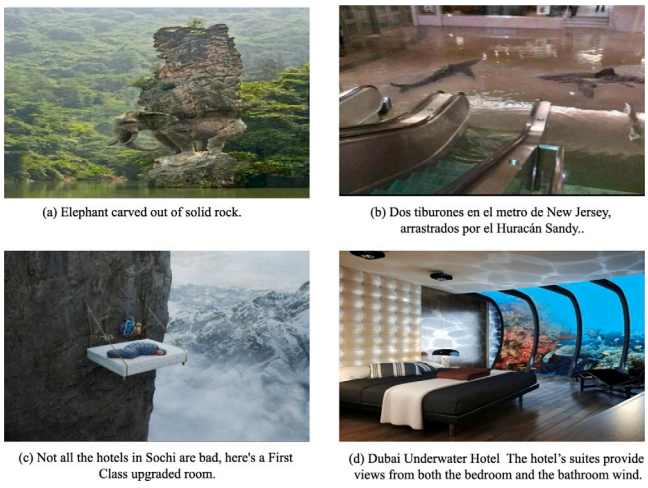

Fig. 1. Some fake news from Twitter. (a) The visual does not provide significant details, and the written content hints that it could be a fake news. (b) Both the text and the visual depiction strongly suggest that it is probably a fake news. (c) An irrelevant news: the text describes a luxury hotel room, but the image depicts a scene of coldness. (d) The fabricated image indicates that it is likely to be fake.

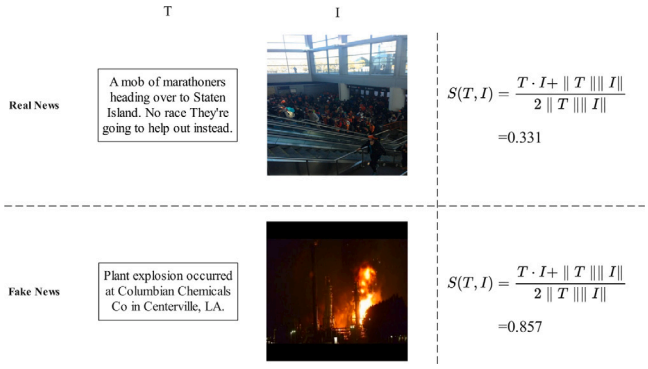

Fig. 2. Regarding concerns about the text–image consistency: a weak correlation between texts and images in some real news, where feature fusion may cause semantic interference between them. Therefore, considering the authenticity of single-modal information (i.e., text or image) separately may be more helpful in accurately detecting fake news. $S(T, I)$ in the figure represents the cosine similarity score between texts and images.

full consideration of single-modal features. Even when the correlation between texts and images is weak, independent single-modal features are still the key decision-making conditions, as is shown in Fig. 2.

Taking the above into consideration, we propose CAMFND, a cross-modal adaptive-aware learning framework for multi-modal fake news detection. We utilize a triple-branch encoder to construct textual, visual, and auxiliary features, and cross-attention is then employed to achieve cross-modal alignment within a unified space. Then we propose a gate fusing and matching mechanism to finish cross-modal interaction, where the gate value could adaptively adjust the degree of multi-modal aggregation, thus reducing the semantic gap and capturing the semantic correlation in terms of text–image consistency. After that, a matching score is computed by the MLP layer, which acts as the weight of choosing multi-modal features or unimodal features. Therefore, a cross-modal adaptive-selective module is proposed to make these features jointly participate in the decision-making process, i.e. relying on unimodal features when cross-modal ambiguity is weak and referring to cross-modal correlations when cross-modal ambiguity is strong. Experimental studies on two widely used datasets (Twitter and Weibo) demonstrate that CAMFND can outperform the state-of-the-art fake news detection methods.

The main contributions of this paper are summarized as follows:

- We design a gated fusing and matching unit, which uses a gated value to choose features adaptively, thereby effectively controlling the degree of fusion of text and image and obtaining multi-modal features.
- We propose a similarity-guided multimodal fake news detection framework, which automatically calculates the similarity score between images and texts, and uses it as the adaptive weight of unimodal and multimodal features to guide joint decision-making to judge the authenticity of news.
- We design an adaptive feature selection module that effectively identifies and selects the features with the strongest representative.

2. Related work

2.1. Multimodal fake news detection

The key to multimodal fake news detection lies in the adjustment of recognition language and visual representations. The attRNN proposed by Jin et al. [4] integrated the textual features, the visual features, and the social context features using attention mechanisms, and finally stitched together these features for classification. Wang et al. [6] introduced EANN, which utilizes the ancillary task of event classification to enhance versatility. Khattar et al. [13] proposed MVAE, a method that trains a decoder from fused features to reconstruct the original text and underlying image features. Singhal et al. [14] proposed a Spotfake, which used VGG and BERT to extract image and text features respectively, and then connect them to justify the content. In recent years, some new fusion technologies have been introduced to replace the original simple splicing technology. The MKEMN method proposed by Zhang et al. [5] learned the multimodal representation of each post, and combined the aligned embeddings of texts, images, and knowledge to detect multimodal fake news. Wu et al. [15] proposed MCAN, which fuses multimodal features by stacking multiple common attention layers, allowing the learning of inter-dependencies between multiple modalities. Wei et al. [16] introduced CMC, a network that trains two unimodal networks through contrastive learning to learn cross-modal associations. Pan et al. [17] proposed MFAE, which firstly uses an improved dynamic routing algorithm to extract more comprehensive semantic visual entity features and utilizes a graph matching network to capture the correspondences between entities within modalities. Jing et al. [18] proposed MPFN, which captures the representational information of each modality at different levels and achieves modality fusion at the same level and across different levels through a mixer. Hua et al. [19] introduced TTEC, which uses the back-translation strategy to increase the size of the dataset, as well as the application of contrastive learning and the multi-head attention mechanism for joint learning. The above work primarily adopts a feature-based fusion approach, thereby weakening the semantic relationships between different modalities. These methods may not fully capture the subtle interactions and contextual dependencies that are vital for a comprehensive understanding of multimodal data. Consequently, the semantic coherence and interpretability of the fused information may be compromised. The above work primarily employs feature-based fusion, thereby weakening the semantic relationship between modalities.

2.2. Cross-modal learning

Nowadays, the concept of semantics has been introduced into fake news detection, and a new methodology has been implemented through the use of some cross-modal similarity techniques. Zhou et al. [9] proposed SAFE to detect fake news by inputting the correlation between textual and visual information in the classifier. However, cross-modal

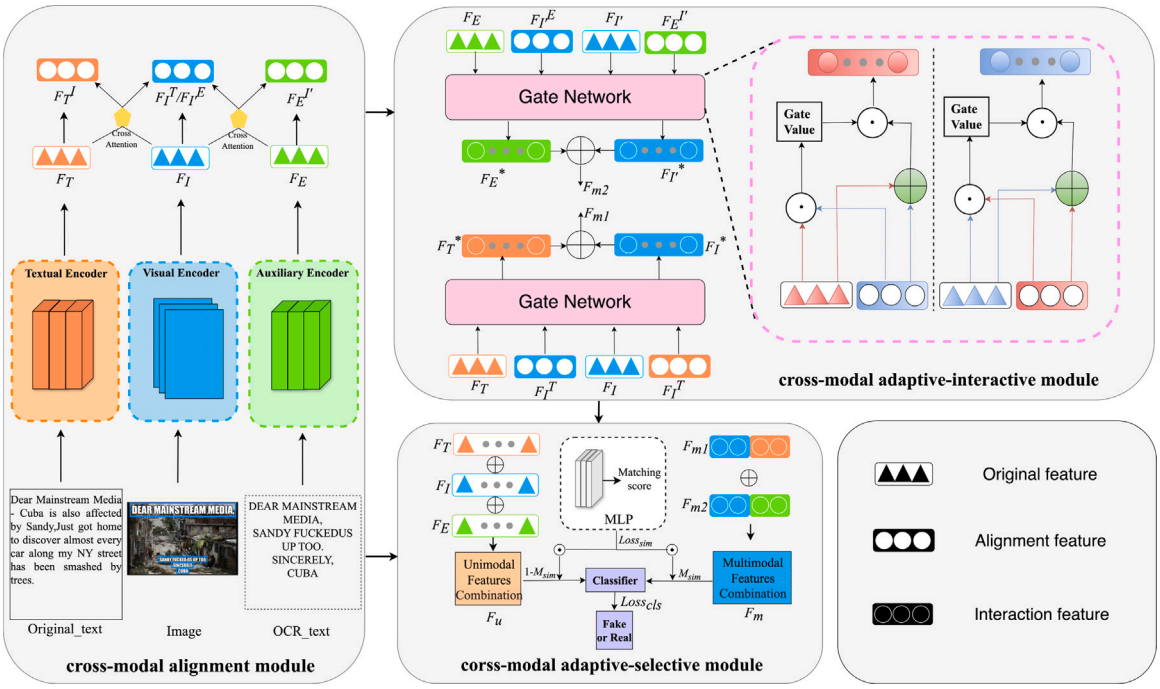

Fig. 3. The framework of CAMFND consists of three main components: (1) A cross-modal alignment module; (2) A cross-modal adaptive-interactive module; (3) A cross-modal adaptive-selective module.

consistency may be limited by the semantic gap in image-to-text translation. Xue et al. [10] introduced MCNN to integrate text semantic features, visual tampering features, and the similarity of text information with visual information to achieve fake news detection. However, the polymerization process which simply connects all the features does not take into account the problem of cross-modal heterogeneity. Wang et al. [20] proposed COOLANT, a cross-modal contrastive learning framework that compares feature representations between different modalities to learn a shared feature space. Li et al. [21] designed CFFN, which introduces cross-modal consistency learning and a fine-grained fusion network to capture and integrate the complex relationships between different modalities. Chen et al. [22] proposed CAFE, which first compresses image and text representations by a variational auto-encoder (VAE) and compares news with correct image–text correspondences by minimizing the Kullback–Leibler (KL) scatter. Subsequently, the KL scatter score was utilized to measure cross-modal consistency. However, due to data limitations, it was difficult to ensure complete alignment of multimodal features. CAFE handled unimodal features by using a variational self-encoder with unshared weights, but this also made it difficult to determine whether the KL scatter is reliable enough to measure cross-modal consistency. Moreover, CAFE imposed a severe penalty on unimodal features when consistency was high, which may have had implications in some news. Based on the problems identified in the above related works, this paper proposes a model that simultaneously addresses the hidden consistency between multiple modalities and their inherent heterogeneity, and effectively solves the problem of engaging more fully with unimodal and multimodal decision-making.

3. Method

In this section, we elaborate on our proposed cross-modal adaptive-aware learning for multimodal fake news detection. The overall framework of the method is shown in Fig. 3.

3.1. Cross-modal alignment module

3.1.1. Text encoder

Given a text x^t with a set of words, we adopt a pretrained modified BERT model to obtain its embedding F_T. Specifically, we accurately reflect the importance of each position in the input sequence by adding a self-attention mechanism [23] to the top layer of the BERT model, instead of simply averaging the output across each position [24].

3.1.2. Image encoder

Given an image x^v with a set of patches, we adopt a pretrained modified ResNet model to obtain its embedding F_I. ResNet can extract more salient features with the help of CBAM [25], which links channel attention with spatial attention, thus the final visual features are calculated incorporating with channel attention and spatial attention.

3.1.3. Auxiliary encoder

It is indicated that textual clues from the attached images present more plentiful and comprehensive contents [26]. Therefore, we consider the text embedded within the image as one of the crucial determinants to integrate a multi-modal encoder. Specifically, given an image x^t with embedding text x^e, we use the pre-trained PP-OCRv3 model to extract the embedded features. A text encoder is utilized correspondingly in the same way as before, thus creating the auxiliary embedding F_E.

3.1.4. Feature alignment

Features from different modalities may have huge semantic gaps so that we need to align the features from different modalities by transforming the unimodal embeddings into a shared space. In this paper, we design a cross attention network to migrate semantic gaps among different modalities. Specifically, the textual feature is mapped as both the Key vector K and the Value vector V, while the visual feature is mapped through three different fully connected layers to serve as the Query vector Q by means of cross-attention mechanism [27], thus forming the enhanced textual feature F_T^I. Similarly, the enhanced visual feature F_I^T and F_I^E modified by textual feature and auxiliary feature respectively, as well as the enhanced auxiliary feature F_E^I modified by the visual feature is computed in the same manner.

3.2. Cross-modal adaptive-interactive module

The multimodal semantic fusion aims to interact among different modalities in fine-grained descriptions. The fine-grained semantic interactions between texts and images, including the fine-grained interactions between different words in the textual sentences and the corresponding image information, as well as the fine-grained interactions between different regions in the visual images and the corresponding text information. Even though the cross-modal alignment module adopts cross attention to interact fundamentally, the issues of information granularity differences between texts and images and semantic matching ambiguity still persist. In this paper, we propose a cross-modal adaptive-interactive module by means of the gated mechanism and selectively integrate the interaction process of texts and images using a multi-modal gated semantic fusion.

The core idea of the adaptive-interactive module is to alleviate semantic confusion caused by information fusion by adjusting the degree of integration. When integrating information from another modality, the flexible gating units in the gating module can adaptively decide which information should be fused. Specifically, for negative sample pairs, the flexible gating units should weaken the fused information from the other modality, ensuring that the fused semantics approximate the original single-modality semantics. Additionally, these gating units can filter out background information that is not important to the matching task, thereby reducing the impact of global semantic fusion on model accuracy. The concrete procedures are presented as below.

The gating units will integrate the fused image features with the original text features while concurrently blending the fused text features with the original image features. Take the base textual features F_T and visual features F_{IT} as an example. Define G_{TI} as the extent of the underlying textual features guided by the visual features. The process can be concluded as Eq. (1) :

$$G_{TI} = \sigma(\sum_{i=1}^{N}(F_T^i \odot F_{IT}^i)) \tag{1}$$

where N is the dimension of the aligned features, $\odot$ denotes element-by-element multiplication, and σ denotes the Sigmoid function.

Through this process, texts and images with higher gating values can effectively facilitate the fusion operation, thereby strengthening the correlation between them. Conversely, for mismatched text and images, the gating value may be reduced to hinder the fusion operation. Subsequently, the final multimodal fusion feature, denoted as F_T^*, is obtained by performing a weighted fusion of features F_T and F_{IT} based on the gating value G_{TI}. This process is illustrated in Eq. (2) :

$$F_T^* = f_t((F_T \oplus F_{IT}) \odot G_{TI}) \tag{2}$$

where f_t denotes the fully connected activation function, $\oplus$ denotes the concatenation symbol, $\odot$ denotes the element-to-element multiplication. Adaptive fusion is reflected in the following aspects: for positive samples, where the correlation between texts and images is high, the value of G_{TI} is larger, resulting in an enhancement of the underlying text features embedded in the image features. Conversely, for negative samples, where there is no significant semantic relationship between the two modalities, the value of G_{TI} is smaller, leading to the suppression of the underlying text features embedded in the image features.

A similar way is repeated to generate F_I^*, thus forming the fusing vector F_{m1}. For the visual features F_I and auxiliary features F_E, we can also obtain the fusing vector F_{m2}.

3.3. Cross-modal adaptive-selective module

In this section, a shared embedding layer consisting of fully connected layers is used to calculate the similarity score. The shared embedding layer network can automatically calculate multi-modal feature information, thereby reducing the feature loss that may be caused by cosine similarity calculation. First, the similarity matching vector $M(T, I)$ of text and image is calculated. The process is shown in:

$$M(T, I) = \sigma(\mathrm{MLP}(F_T^* \oplus F_I^*)) \tag{3}$$

where MLP is a multilayer perceptron structure, including a linear layer from the input dimension to the hidden dimension, followed by a ReLU activation function and a Dropout layer. Then, there is a linear layer that maps the hidden layer output, and finally, the output is passed through the Sigmoid activation function, mapping to the range of 0 to 1, characterizing the matching degree of texts and images. Similarly, the similarity matching vector $M(E, I)$ between images and embeddings can be obtained as well. Finally, the final similarity matching score $M(sim)$ is obtained by averaging the two values:

$$M(sim) = (M(T, I) + M(E, I))/2 \tag{4}$$

In this paper, we propose a cross-modal adaptive-selective module weighted by the matching score, between the uni-modal features and the multi-modal features. When the matching score between the texts and images of a news article is high, it indicates strong consistency and complementarity between them. In such cases, the multimodal feature combination is given a higher weight, allowing it to play a more significant role in the decision-making process. Conversely, if the matching score between the texts and images is low, suggesting some degree of inconsistency or contradiction, this may indicate potential fake news. At this point, the multimodal feature combination might contain noise or interference factors, and its influence in the decision-making process is reduced.

Therefore, we use the similarity matching score $M(sim)$ as the weight of the multi-modal feature combination F_m, and concatenate it with the uni-modal feature combination F_u to complement each other's information:

$$F_m = (F_T^* \oplus F_I^* \oplus F_E^* \oplus F_{I'}^*) \quad F_u = F_T \oplus F_I \oplus F_E \tag{5}$$

The likelihood that a news article is fake is then predicted by feeding a combination of these features into a fake news detector. The process is shown below:

$$C = \mathrm{SoftMax}\{W_c[(M(sim) \odot F_m) \oplus ((1 - M(sim)) \odot F_u)] + B_c\} \tag{6}$$

where C represents the probability of news category prediction, W_c and B_c are the weight parameter matrix and offset term respectively.

3.4. Model learning

In order to make the probability of a news article being forged as close as possible to its true label value, a loss function based on cross-entropy is defined as:

$$\mathcal{L}_{cls} = -\mathbb{E}_{(a,y)\sim(A,Y)} \left[y \cdot \log C + (1 - y) \cdot \log(1 - C)\right] \tag{7}$$

where Y is the true and false label category set of news, A is the news set, and y is the actual label of news a.

On the other hand, expressed from the perspective of similarity, the loss function is designed to motivate the model to better learn the matching relationship between texts and images, which is expressed by

$$\mathcal{L}_{\text{sim}} = -\mathbb{E}_{(a,y)\sim(A,Y)}\{y \cdot \log[1 - M(\text{sim})] + (1 - y) \cdot \log M(\text{sim})\} \tag{8}$$

Overall, the loss function consists of two parts: classification loss and similarity loss. These two parts of the loss respectively measure the model's performance in predicting the authenticity of news, and the accuracy of the model's assessment of the similarity between texts and images. The two parts of the loss are linearly weighted and added, where α and β represent the weights of the two loss functions respectively. In this way, a loss value that comprehensively reflects the performance of the model can be obtained, allowing both aspects to be

Table 1
Comparison of experimental results for methods.

Dataset	Method	Accuracy	Real information			Fake information		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	att-RNN [4]	0.788	0.738	0.890	0.807	0.862	0.686	0.764
	EANN [6]	0.816	0.810	0.810	0.810	0.820	0.820	0.820
	MVAE [13]	0.824	0.802	0.875	0.837	0.854	0.769	0.809
	Spotfake [14]	0.892	0.847	0.656	0.739	0.902	**0.964**	**0.932**
	MKEMN [5]	0.824	0.723	0.819	0.798	0.823	0.799	0.812
	SAFE [9]	0.816	0.816	0.818	0.817	0.818	0.815	0.817
	MCNN [10]	0.823	0.787	0.848	0.816	0.858	0.801	0.828
	MCAN [15]	0.899	0.884	0.909	0.897	0.913	0.889	0.901
	CAFE [22]	0.840	0.825	0.851	0.837	0.855	0.830	0.842
	LIIMR [12]	0.900	0.882	0.823	0.847	0.908	0.941	0.925
	MFIR [28]	**0.925**	**0.906**	0.848	0.876	0.901	0.839	0.869
	CAMFND	0.908	0.892	**0.934**	**0.912**	**0.918**	0.910	0.914
Twitter	att-RNN [4]	0.682	0.603	0.770	0.676	0.780	0.615	0.689
	EANN [6]	0.719	0.771	0.870	0.817	0.642	0.474	0.545
	MVAE [13]	0.745	0.689	0.777	0.730	0.801	0.719	0.758
	Spotfake [14]	0.777	0.832	0.606	0.701	0.751	0.900	0.820
	MKEMN [5]	0.714	0.634	0.814	0.831	0.814	0.756	0.708
	SAFE [9]	0.762	0.695	0.811	0.748	0.831	0.724	0.774
	MCNN [10]	0.784	0.790	0.787	0.788	0.778	0.781	0.779
	MCAN [15]	0.809	0.732	0.871	0.795	0.889	0.765	0.822
	CAFE [22]	0.806	0.805	0.813	0.809	0.807	0.799	0.803
	LIIMR [12]	0.831	0.836	0.832	0.830	0.825	0.830	0.827
	MFIR [28]	0.858	0.867	0.871	0.869	0.850	0.848	0.849
	CAMFND	**0.934**	**0.931**	**0.934**	**0.933**	**0.952**	**0.942**	**0.947**

optimized simultaneously during the training process. By minimizing the loss function, the optimal parameter $\hat{\theta}^*_{mod}$ can be obtained:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{sim} \tag{9}$$

$$\left(\hat{\theta}^*_{\text{mod}}\right) = \arg\min_{\theta^*_{\text{mod}}} \mathcal{L}_{\text{total}} \tag{10}$$

4. Experiments

In this section, we evaluate our proposed CAMFND model by comparing it with several benchmark models. Subsequently, ablation and visual experiments are conducted to validate the effectiveness of each module.

4.1. Dataset

The datasets are described as follows:

(1) Weibo dataset: was released by Jin et al. [29], which has been widely used in prior multimodal fake news detection works. The real ones were collected from Xinhua News Agency, an authoritative news source in China. The fake ones were gathered by crawling the official fake news debunking system of Weibo over a time span from May 2012 to January 2016. In experiments, the training set contains 7532 news, including 3749 fake news and 3783 non-fake news; the test set contains 1996 posts.

(2) Twitter dataset: was released for MediaEval Verifying Multimedia Use task [30]. Given the focus on the text and image, following existing works, we filter the tweets with videos attached. In experiments, we keep the same data split scheme as the benchmark [30], which is also the same as all the compared methods. The training set contains 6840 real news and 5007 fake news and the test set contains 1406 posts.

4.2. Experimental setup

In the cross-modal alignment module, the dimensionality of the text features obtained from the textual encoder is 768. The dimensionality of the visual features obtained from the visual encoder is 2048. The final cross-modal features obtained are all of 32×32 dimensions. In the cross-modal adaptive-interactive module, the dimensions of the gated values are 32, thus the dimensions of the adaptive fusion features subsequently updated with this gated value are all 32×64. In the cross-modal adaptive-selective module, the weight of classification loss is set as 0.6 and that of similarity loss is equal to 0.4. Moreover, Adam is chosen as the optimizer. The batch size is 32 and the number of epochs is 100. The learning rate for Weibo is set to 0.001 while the learning rate for Twitter is set to 0.0001.

4.3. Performance comparison

Table 1 compares the performance of our model with the baseline model. On the two real datasets, our model achieves the highest accuracy of 90.8% and 93.4%, respectively. Notably, our method exhibits evident superiority in the English Twitter dataset.

From the perspective of semantics, our algorithm exceeds att-RNN, EANN, MVAE, Spotfake, and MKEMN a lot, which indicates the importance of image–text semantic information. On the other hand, there is no cross-modal interaction between SAFE and MCNN, leading to the problem of modal heterogeneity. The experimental results on the Weibo dataset show that the accuracy of our model is 9.2% higher than that of SAFE and 8.5% higher than that of MCNN, which proves that our model makes full use of multimodal interaction data to enrich the correlation characteristics between texts and images. Also, the superiority over MCAN, CAFE, and LIIMR by 0.9%, 6.8%, and 0.8% respectively validates the effectiveness of adaptive learning of cross-modal tasks.

4.4. Ablation study

We conducted a series of ablation experiments to evaluate the effectiveness of each component of our approach. The results are shown in Figs. 4 and 5.

From the results shown in Fig. 4, we have the following observations: (1) Compared to CAMFND, the alignment+interactive+selective approach performs weaker, indicating that the cross-modal features learned by the proposed gated fusion network are more effective than simply concatenating unimodal features as cross-modal features. (2) The alignment+interactive approach performs poorly, proving that adaptively aggregating complementary unimodal representations and cross-modal correlations can improve the accuracy of the model. (3) Compared to CAMFND, the alignment approach performs

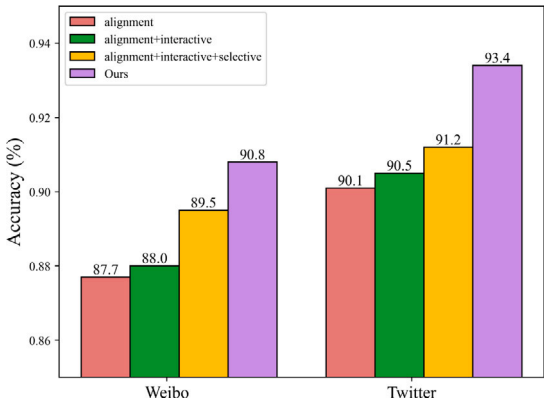

Fig. 4. CAMFND ablativity experiments.

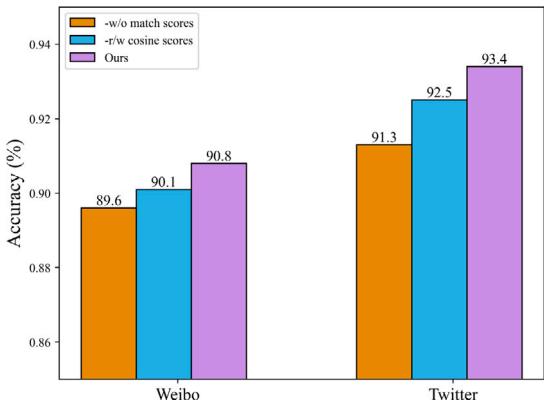

Fig. 5. CAMFND ablativity experiments.

worse,
demonstrating that designing cross-modal adaptive-interactive module and adaptive-selective modules are crucial for addressing semantic correlation and consistency. In Fig. 5, both w/o match score and r/w cosine scores perform worse than CAMFND, indicating the effectiveness of the similarity matching network designed in the model.

- alignment: Remove the second and third modules, and only reserve the ablation experimental results of multimodal decision-making obtained by the cross-modal alignment module.
- alignment+interactive: Remove the third module, and only reserve the ablation experimental results of multimodal decision-making obtained by gated fusion.
- alignment+interactive+selective: Remove the gated fusion and reserve the ablation experimental results of multimodal features and decision-making are obtained by cross-modal interaction module.
- -w/o *match scores*: After removing the similarity score obtained by MLP, there are only ablation experimental results of uni-modal and multimodal joint decision-making.
- -r/w *cosine scores*: In this group of experiments, the similarity score is calculated by the traditional cosine similarity.
- *Ours*: The complete model architecture proposed in this article.

4.5. Case study

This section studies the news in the dataset and compares its true label Y with the similarity score S obtained by the model. Some examples are shown in Figs. 6 and 7. The following can be observed: (1) In some cases, there are differences between texts and images. For example, as shown in Fig. 6(b), the news event described in the text does not receive particularly relevant support in the image. (2) The model proposed in this paper helps to accurately evaluate the similarity between news textual and visual information.

For example, the real news shown in Fig. 6(a) has a high similarity score, and the model in this paper correctly labels them as real news;

Fig. 6. Case Study 1.

Fig. 7. Case Study 2.

while the fake news shown in Fig. 6(b) has a low similarity score, and the model in this paper correctly labels them as fake news. The same rule applies for Fig. 7(a) and (b), which indicates how the text–image consistency adaptively influence of degree of authenticity of a news.

5. Conclusion

In this paper, we propose a cross-modal adaptive-aware learning framework for multi-modal fake news detection, aiming to reduce semantic ambiguities among different modalities, which has broad application potential in future social media. Our proposed CAMFND consists of three parts: a cross-modal alignment module to transform the heterogeneous unimodality features into a shared semantic space, a cross-modal adaptive-interactive module to capture the semantic correlation and consistency, computed by the multi-modal gated fusion unit, and a cross-modal adaptive-selective module to decide the semantic meaning or bias, guided by the multi-modal semantic matching score. CAMFND enhances the fake news detection by intelligently and dynamically combining features from uni-modality and identifying correlations across different modalities. It leverages unimodal features in scenarios with low cross-modal ambiguity, while utilizing cross-modal correlations in cases of high cross-modal uncertainty. The experimental results show that CAMFND significantly surpasses prior methodologies and sets new benchmarks on both datasets, marking a notable advancement in performance.

CRediT authorship contribution statement

Ying Guo: Writing – original draft, Supervision, Resources, Methodology, Formal analysis, Conceptualization. **Yuan Li:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Kexin Zhen:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Bingxin Li:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Jie Liu:** Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the MOE (Ministry of Education in China) Liberal arts and Social Sciences Foundation (24YJC86007), the Key Supported Program of Joint Fund of the National Natural Science Foundation of China (U23B2029), National Natural Science Foundation of China (62076167) and the Yuxiu Innovation Project of NCUT (2024NCUTYXCX102).

Data availability

Data will be made available on request.

References

[1] T. Mitra, G.P. Wright, E. Gilbert, A parsimonious language model of social media credibility across disparate events, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 126–145.
[2] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: The Web Conference, 2011.
[3] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, 2016.
[4] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: The 2017 ACM, 2017.
[5] H. Zhang, Q. Fang, S. Qian, C. Xu, Multi-modal knowledge-aware event memory network for social media rumor detection, in: The 27th ACM International Conference, 2019.
[6] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: ACM SIGKDD, 2018, pp. 849–857.
[7] Y.J. Lu, C.T. Li, GCAN: Graph-aware co-attention networks for explainable fake news detection on social media, 2020.
[8] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Signal Process. 14 (3) (2020) 478–493.
[9] X. Zhou, J. Wu, R. Zafarani, : Similarity-aware multi-modal fake news detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2020, pp. 354–367.
[10] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, Detecting fake news by exploring the consistency of multimodal data, Inf. Process. Manage. 58 (5) (2021) 102610.
[11] H. Choi, Y. Ko, Effective fake news video detection using domain knowledge and multimodal data fusion on youtube, Pattern Recognit. Lett. 154 (2022) 44–52.
[12] S. Singhal, T. Pandey, S. Mrig, R.R. Shah, P. Kumaraguru, Leveraging intra and inter modality relationship for multimodal fake news detection, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 726–734.
[13] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.
[14] S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, S. 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.
[15] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of ACL 2021, 2021.
[16] Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, D. Li, Cross-modal knowledge distillation in multi-modal fake news detection, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 4733–4737.
[17] Z. Pan, Y. Mao, L. Xiong, T. Pang, P. Ping, MFAE: Multimodal fusion and alignment for entity-level disinformation detection, Pattern Recognit. Lett. (2024).
[18] J. Jing, H. Wu, J. Sun, X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, Inf. Process. Manage. 60 (1) (2023) 103120.
[19] J. Hua, X. Cui, X. Li, K. Tang, P. Zhu, Multimodal fake news detection through data augmentation-based contrastive learning, Appl. Soft Comput. 136 (2023) 110125.
[20] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, S. Wang, Cross-modal contrastive learning for multimodal fake news detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5696–5704.
[21] J. Li, Y. Bin, J. Zou, J. Wei, G. Wang, Y. Yang, Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection, in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, 2023, pp. 1–7.
[22] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2897–2905.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, ArXiv.
[24] Z. Lin, M. Feng, C.N.d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, 2017, arXiv preprint arXiv:1703.03130.
[25] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: ECCV, 2018, pp. 3–19.
[26] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, Y. 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, 2021, pp. 1212–1220.
[27] M. Gheini, X. Ren, J. May, Cross-attention is all you need: Adapting pretrained transformers for machine translation, 2021, arXiv preprint arXiv:2104.08771.
[28] L. Wu, Y. Long, C. Gao, Z. Wang, Y. Zhang, MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection, Inf. Fusion 100 (2023) 101944.
[29] Z. Jin, J. Cao, Y. Zhang, J. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification, IEEE Trans. Multimed. 19 (3) (2016) 598–608.
[30] Z. Jin, J. Cao, Y. Zhang, Y. Zhang, MCG-ICT at MediaEval 2015: Verifying multimedia use with a two-level classification model, in: MediaEval, 2015.