Multi-Task Network Guided Multimodal Fusion for Fake News Detection

Jinke Ma[†] 2231968@s.hlju.edu.cn
Liyuan Zhang[†] 2231976@s.hlju.edu.cn
Yong Liu[*] 2010023@hlju.edu.cn
Wei Zhang[*] zhangwei_jsj@hlju.edu.cn
Heilongjiang University, Harbin, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Fake news detection has become a popular research topic in the multimodal field. Existing multimodal fake news detection research utilizes a series of feature fusion networks to collect useful information from news posts of different modalities. However, how to form effective cross-modal interaction features? How to utilize cross-modal correlations to facilitate cross-modal interactions? These are still open questions. In this paper, we introduce MMFND, a multi-task network guided multimodal fusion framework for fake news detection, which introduces multi-task network for feature refinement and fusion. Paired CLIP encoders are used to extract modality-aligned depth representations that enable accurate measurement of cross-modal correlations. Cross-modal interaction features are weighted using normalized cross-modal correlations to enhance feature fusion. Extensive experiments on typical fake news datasets show that MMFND outperforms state-of-the-art methods.

Keywords: Multi-task Networks, Cross-modal Correlation, Feature Refinement

1. Introduction

The rise of Online Social Networks (OSNs) has made information sharing easier, but also facilitates the spread of fake news Zubiaga et al. (2018). As a result, there is a growing need to monitor the credibility of online posts and a growing interest in automated fake news detection techniques Sharma et al. (2019). Early research focused on textual or visual content alone, using decision tree classifiers Liu et al. (2015) or convolutional neural networks Yu et al. (2017), but this is insufficient as most social media posts contain multimodal information.

In multimodal fake news detection, some methods aggregate multimodal features Wang et al. (2023); Wu et al. (2021); Zhu et al. (2024). Some methods use internet evidence to fact-check image-title pairings Abdelnabi et al. (2022). These approaches often fail to take full advantage of the correlation between different modal features and fail to generate efficient cross-modal interaction features.

To address these issues, this paper introduces MMFND, a multi-task network guided multimodal fusion framework for fake news detection, combining the Progressive Layered Extractor (PLE) Tang et al. (2020) with the CLIP model Radford et al. (2021). Figure 1 illustrates four examples of using MMFND for fake news detection, demonstrating MMFND's ability to reduce focus on multimodal features when correlations between different modalities are weak, thereby flexibly aggregating information.

[†] These authors contributed equally to this work and should be considered co-first authors.
[*] Corresponding author.

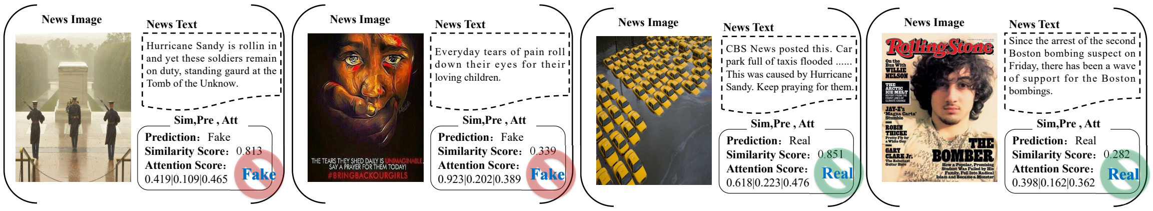

News Image
News Text
Hurricane Sandy is rollin in and yet these soldiers remain on duty, standing gaurd at the Tomb of the Unknow.
Sim,Pre , Att
Prediction: Fake
Similarity Score: 0.813
Attention Score: 0.419|0.109|0.465
Fake
News Image
THE TEARS THEY SHED DAILY IS UNIMAGINABLE. SAY A PRAYER FOR THEM TODAY!
#BRINGBACKOURGIRLS
News Text
Everyday tears of pain roll down their eyes for their loving children.
Sim,Pre , Att
Prediction: Fake
Similarity Score: 0.339
Attention Score: 0.923|0.202|0.389
Fake
News Image
News Text
CBS News posted this. Car park full of taxis flooded ...... This was caused by Hurricane Sandy. Keep praying for them.
Sim,Pre , Att
Prediction: Real
Similarity Score: 0.851
Attention Score: 0.618|0.223|0.476
Real
News Image
Rolling Stone
THE BOMBER
News Text
Since the arrest of the second Boston bombing suspect on Friday, there has been a wave of support for the Boston bombings.
Sim,Pre , Att
Prediction: Real
Similarity Score: 0.282
Attention Score: 0.398|0.162|0.362
Real

Figure 1: Examples of using MMFND to detect fake news. For each news item, three attention scores are provided: text score, image score, and fusion score.

The contributions of this paper are threefold,

• Proposed for a new framework MMFND, which provides a new approach for fake news detection task.

• The large-scale visual language model CLIP is proposed to extract aligned modal embeddings and further refine the features by combining with a multi-task network to form representative and efficient features. New ideas are provided on the processing of features.

• Extensive experiments are conducted on two well-known datasets, and the results show that MMFND outperforms state-of-the-art fake news detection methods. In addition, the ablation study validates the effectiveness of combining multitasking networks and extracting aligned modal embeddings. [1].

2. Related Work

2.1. Unimodal Fake News Detection

Early work on fake news detection typically relied solely on visual or textual information from news posts. For text, some studies utilized lexical and semantic properties to detect fake news (Granmo et al. (2019)), Allein et al. (2021) investigated the correlation between user-generated content and shared content , and some used Transformers (Vaswani et al. (2017)) to capture lexical and semantic attributes (Bhattarai et al. (2021)). Some studies(Ma et al. (2023a,b)) have also employed contrastive learning models to capture features in the propagation path. For images, Jin et al. (2016) discovered significant differences in the image distributions between real and fake news. Cao et al. (2020) examined forensic features, semantic features, statistical features, and contextual features of images. With the development of image generation technology, image tampering detection has also become important (Chen et al. (2021)). Cao et al. (2018) provided a comprehensive summary and introduction to automatic rumor detection technologies. However, these methods often overlooked the complementary nature of information across different modalities, thereby limiting model performance.

2.2. Multimodal Fake News Detection

The key to multimodal fake news detection lies in how to facilitate cross-modal interaction and coordinate text and visual information to form cross-modal features. Singhal et al.

[1] https://github.com/diga7654321/MMFND

(2020) integrated pretrained XLNet and ResNet to extract features, Wang et al. (2018) introduced an additional discriminator to mitigate the impact of specific events, and Zheng et al. (2022) proposed a multimodal attention network to fuse textual, visual, and social graph features. Although these methods have made some progress in multimodal fake news detection (FND), forming effective cross-modal features and efficiently utilizing the correlations between visual and textual content in news remain significant challenges.

2.3. Pre-training model

Transformers have been proven to be an effective Pretrained Model (PTM) architecture (Vaswani et al. (2017)). Devlin et al. (2018) introduced BERT, which has been widely applied to text feature extraction tasks due to its exceptional text feature extraction capabilities. ResNet (He et al. (2016)) introduced residual structures to enhance network depth and improve learning capabilities, and these pretrained models have achieved outstanding performance in many downstream tasks.

Recently, large-scale pretrained models have demonstrated their significance in advancing research in the multimodal domain (Han et al. (2021)). For instance, CLIP (Radford et al. (2021)) achieves effective alignment of textual and visual features through contrastive learning methods and leveraging large-scale image-text pair datasets. Multimodal learning based on CLIP has been applied to various downstream tasks and has achieved excellent results.

3. PROBLEM STATEMENT

We define the task of multimodal fake news detection as a binary classification problem, focusing on the authenticity of multimodal news on social media. A piece of multimodal news consists of text and images, and the model outputs a label $y = \{0, 1\}$ to indicate the veracity of the multimodal news, where 0 and 1 represent the sample being true and false, respectively.

3.1. Model Overview

Figure 2 illustrates the network design of MMFND. The entire process consists of four main components, (1)Feature Extraction Module, this module extracts deep representations of text and images using BERT (Devlin et al. (2018)), ResNet (He et al. (2016)), and a pair of CLIP encoders (Radford et al. (2021)). (2)Feature Refinement and Fusion Based on PLE, this component treats unimodal representation and cross-modal interaction as two tasks within a single modality. By utilizing shared and specific expert groups within PLE, it extracts characteristic and common features of different modalities to enhance feature fusion. (3)Cross-Modal Correlation Measurement and Re-weighting, this step involves calculating the correlation between text and image features using CLIP to assess the importance of fused features. An attention mechanism is used to adaptively weight and fuse the features from different modalities. (4)Classifier and Objective Function, the objective function of MMFND is to minimize cross-entropy loss to correctly predict true and false news. A two-layer fully connected network is used as the classifier.

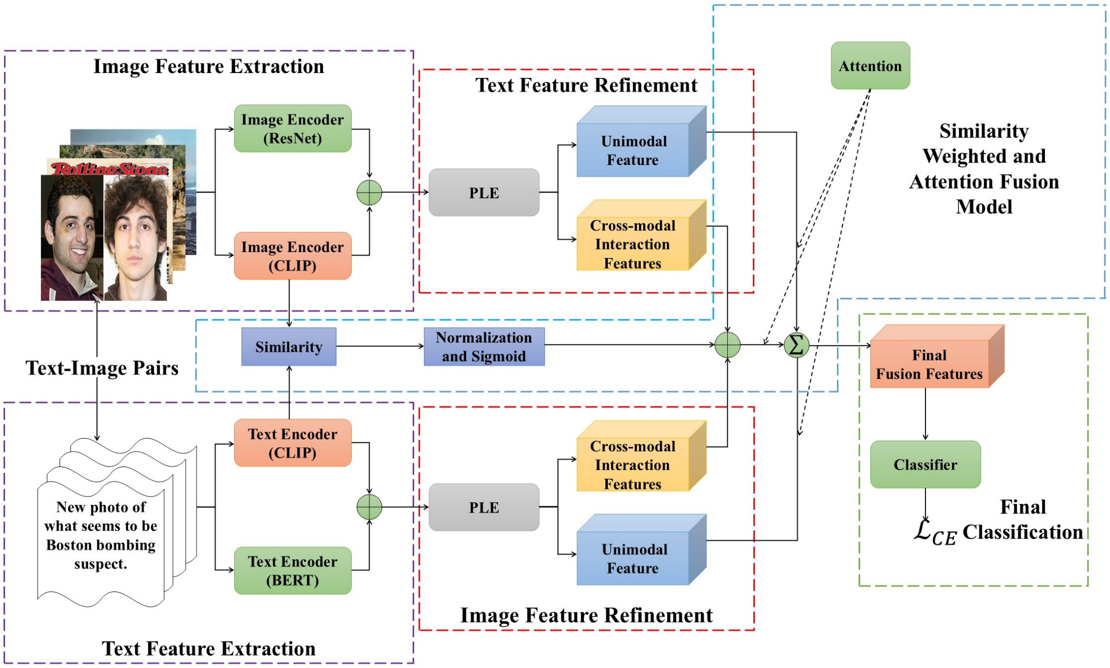


Figure 2: The architecture of the proposed MMFND method.

3.1.1. Feature Extraction

Assume the input multimodal news is represented as $\mathcal{N} = [\boldsymbol{T}, \boldsymbol{V}] \in \mathcal{D}$, where $\boldsymbol{T}$, $\boldsymbol{V}$, and $\mathcal{D}$ denote the text, image, and dataset, respectively. We use established pretrained models to extract features from the text $\boldsymbol{T}$ and the image $\boldsymbol{V}$ in the posts. These models include BERT (Devlin et al. (2018)), ResNet (He et al. (2016)), and CLIP (Radford et al. (2021)). Considering that the images in news posts often contain embedded text, we use a public API to extract the embedded text from the images, incorporating it as supplementary text.

We utilize the pretrained BERT model (Devlin et al. (2018)) to encode the features in $\boldsymbol{T}$. The content of $\boldsymbol{T}$ includes both the original text and the extracted embedded text from the images. After encoding $\boldsymbol{T}$ using BERT, we obtain the textual feature $t^b \in \mathbb{R}^{d_b}$, where d_b is the dimension of the text embedding. For the image $\boldsymbol{V}$, we use the pretrained ResNet (He et al. (2016)) to extract the embedded representation, denoted as $v^r \in \mathbb{R}^{d_r}$, where d_r is the dimension of the image embeddings.

Additionally, we use a pair of pretrained CLIP encoders (Radford et al. (2021)) to extract modality-aligned text and visual embeddings t^c and v^c, respectively. To enhance the representation capacity of the unimodal branches, we perform embedding concatenation within the unimodal text and image branches, resulting in two concatenated embedding features, t^s and v^s.

$$\begin{cases} t^s = concat(t^b, t^c) \\ v^s = concat(v^r, v^c) \end{cases} \tag{1}$$

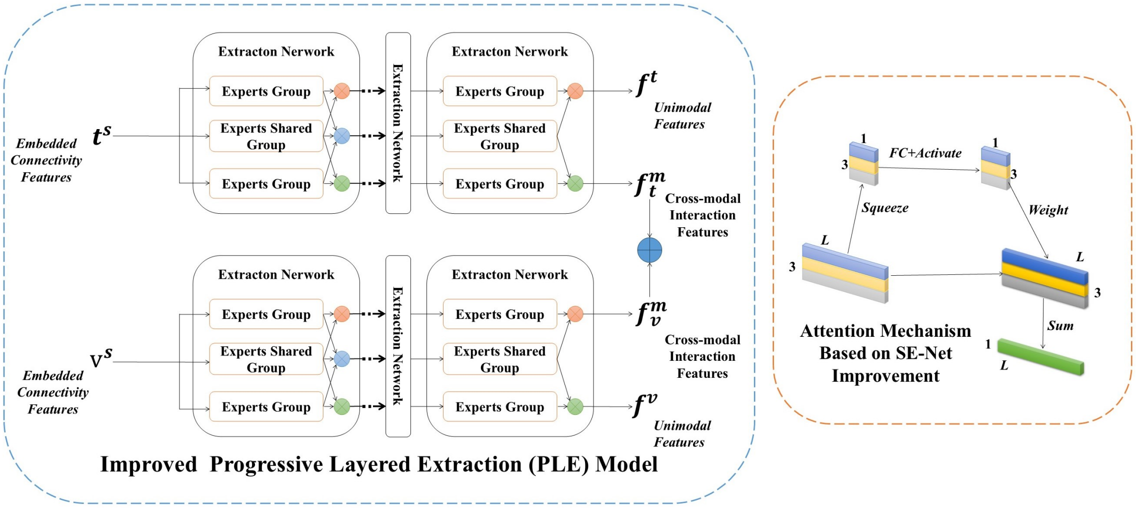


Figure 3: Improved progressive layered extraction (PLE) model and attention mechanism based on SE-Net improvement .

3.1.2. Feature Refinement and Fusion Based on PLE

In the refinement and fusion phase, we introduced the PLE module (Tang et al. (2020)) to refine the embedded connected features t^s and v^s formed in the feature extraction phase. Figure 3 shows our improved PLE model. The design of PLE aimed to isolate from the original features the parts that are beneficial for different tasks. Due to its unique structure, it possesses strong feature refinement capabilities. Specifically, the PLE model comprises multiple extraction networks, including exclusive layers (Experts Group A and B) and shared layers (Experts Shared Group), with each group containing multiple experts. In the PLE model, we treat the mining of unimodal representations and cross-modal interactions as two different sub-tasks within a single modality. For each modality, the two exclusive layers in the PLE form a feature extraction network for a single task with the shared layer, respectively. We denote the feature extraction network for the task of mining unimodal representations as G_u, and the feature extraction network for the task of cross-modal interactions as G_m. We stitch together the features extracted by the feature extraction network for the task of cross-modal interaction to form cross-modal features f^m. Overall, after refinement and fusion, this process results in new representations from three branches, the unimodal features of text f^t and vision (f^v), and the cross-modal feature f^m.

$$\begin{cases} f^t, f^v = G_u(t^s, v^s) \\ f_t^m, f_v^m = G_m(t^s, v^s) \\ f^m = f_t^m + f_v^m \end{cases} \tag{2}$$

Where f_t^m and f_v^m denote the features extracted by the feature extraction network for the task of cross-modal interaction.

3.1.3. Measurement of Cross-modal Correlations and weighting

Cross-modal correlations can affect the effectiveness of cross-modal features. In some news posts, the visual and textual information may have little semantic connection. In such cases, cross-modal features can become noise points for the model, negatively impacting performance. To better adjust the relationships between different modalities, we use CLIP to compute the cosine similarity between text and image features. This similarity is then used to re-weight and adjust the strength of the multimodal features. Specifically, we first calculate the cosine similarity between text features and image features.

$$sim = \frac{t^c \cdot (v^c)^T}{\|t^c\|\|v^c\|} \tag{3}$$

Then, we apply normalization and the Sigmoid function to map the similarity to the [0,1] range.

$$sim_{sigmoid} = \sigma(sim_{normalized}) = \frac{1}{1 + e^{-sim_{normalized}}} \tag{4}$$

The normalization is done by calculating the running mean (μ) and standard deviation (σ) during training, then subtracting the running mean from the similarity value (sim) and dividing it by the running standard deviation. Unlike the contrastive learning paradigm, normalization helps in calculating the similarity without the need to compare the news post with other instances.

$$sim_{normalized} = \frac{sim - \mu}{\sigma} \tag{5}$$

Considering the contrastive learning pretraining method of CLIP, we can assume that the textual and visual representations t^c and v^c extracted by CLIP are aligned. This makes the calculated correlations more accurate and interpretable. Using the obtained cross-modal correlations, we readjust the strength of the cross-modal features.

$$f^{cross} = Sigmoid(Std(sim)) \cdot f^m \tag{6}$$

Finally, we obtain the features from two unimodal branches f^t and f^v, as well as the cross-modal branch feature f^{cross}. We use a modality attention module, adapted from SE-Net (Hu et al. (2018)), to capture dependencies between features for adaptive feature fusion. The attention module we use is illustrated in Figure 3. First, we concatenate f^t, f^v and f^cross into an $L \times 3$ matrix, which is then compressed using average pooling and max pooling to obtain an initial weight vector of 1×3. The initial weight vector is fed into two fully connected layers of size 3×3 and activated using the GELU activation function. The output is then normalized to the [0, 1] range using the Sigmoid function, resulting in the attention weights for each channel, $att = \{att_t, att_v, att_{cross}\}$. Each attention weight att is multiplied by the corresponding features f^t, f^v and f^{cross}. The weighted features are then summed to obtain the aggregated feature m^{Agg}.

$$m_{Agg} = att_t \cdot f^t + att_v \cdot f^v + att_{cross} \cdot f^{cross} \tag{7}$$

3.1.4. Classification and Objective Function

The aggregated representation m_{Agg} is fed into a two-layer fully connected network, which serves as the classifier F_{cls} to predict the label $\hat{y}$. The objective function of MMFND is to minimize cross-entropy loss to accurately predict true and false news.

$$\mathcal{L}_{CE} = y\log(\hat{y}) + (1-y)\log(1-\hat{y}). \tag{8}$$

Specifically, the entire training process of the model is as follows. First, deep representations are extracted from the text and images. Then, these features are refined and fused using PLE. After that, these representations undergo a series of weighting and fusion operations. Finally, the integrated features are further processed to form a single output that approximates the true value.

$$\hat{y} = F_{Cls}\left(F_{Mix}\left(G\left(F_{Txt}\left(X_{Txt}\right)\right), G\left(F_{Img}\left(X_{Img}\right)\right)\right)\right) \tag{9}$$

Where, F_{Txt} and F_{Img} represent the unimodal feature extractors, including BERT, ResNet, and a pair of CLIP encoders. F_{Mix} represents the feature weighting and fusion mechanism, F_{Cls} represents the classification head, and G represents the Progressive Layered Extraction model in the multi-task network.

4. EXPERIMENTS

4.1. Experimental Setup

Table 1: Experimental Results.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	EANN	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	0.858	0.801	0.828	0.787	0.848	0.816
	MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	LIIMR	0.900	0.882	0.823	0.847	0.908	0.941	0.925
	CMC	0.908	**0.940**	0.869	0.899	0.876	**0.945**	0.907
	FND-CLIP	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	MMFND	**0.935**	0.930	**0.941**	**0.935**	**0.940**	0.929	**0.934**
Twitter	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	MCAN	0.809	**0.899**	0.765	0.822	0.732	0.871	0.795
	CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	LIIMR	0.831	0.836	0.832	0.830	0.825	0.830	0.827
	FND-CLIP	0.864	0.833	0.907	0.869	0.899	0.820	0.858
	MMFND	**0.896**	0.892	**0.912**	**0.902**	**0.901**	**0.878**	**0.889**

4.1.1. Dataset

Our model was evaluated on two real-world datasets, Weibo (Nan et al. (2021)) and Twitter (Boididou et al. (2018)). The Weibo dataset, sourced from Xinhua News Agency and Sina Weibo, includes Weibo IDs, text, and images, and is widely used in Chinese fake news detection research. The Twitter dataset comes from Twitter and is used for the MediaEval multimedia verification task, including tweet IDs, text, and images, widely applied in English and cross-domain fake news detection studies. The use of these multilingual datasets enhances the reliability of our findings.

4.1.2. Training Detail

For the text samples, firstly, for the selection of BERT pre-training models, we use the "chinese-roberta-wwm-ext" model, which is a BERT variant optimised for the Chinese data, and the "bert- base-uncased" model on the English dataset, combined with an attention mechanism-based post-processing method (Jawahar et al. (2019)). The maximum length for input text was set to 300 words. For image samples, for ResNet selection, we use the pre-trained ResNet-101 to extract visual features with an input image size of 224×224 pixels. In choosing the CLIP model, we used the modified Chinese-CLIP-ViT for the Chinese dataset and CLIP-ViT-B/32 for the English dataset. During the entire training phase, we fine-tuned ResNet and froze the weights of the BERT and CLIP models. The number of experts in the PLE network was set to 3. The two hidden layers of the classifier were sized 64 and 2, respectively. The learning rate was set at 1×10^{-4}, with a weight decay rate of 0.01. To avoid overfitting, we selected the iteration with the highest test accuracy from 100 rounds of training as the final result.

4.1.3. Baseline Methods

For fair and reproducible comparisons, we primarily used ACC(Accuracy) and F1-score as evaluation metrics. ACC refers to the proportion of samples correctly predicted by the classification model relative to the total number of samples. F1-score is the harmonic mean of precision and recall, mainly used to assess the performance of binary classification models in scenarios where positive and negative samples are imbalanced. We compared the MMFND model against the following strong baselines:

- **EANN**(Wang et al. (2018)), which is a GAN-based model that aims to remove the event-specific features.
- **MVAE**(Khattar et al. (2019)), which uses a variational autoencoder coupled with a binary classifier to learn shared representations of text and image.
- **MKEMN**(Zhang et al. (2019)), which exploits the external knowledge-level connections to detect fake news.
- **SAFE**(Zhou et al. (2020)), which measures cross-modal similarity for fake news detection.
- **MCNN**(Xue et al. (2021)), which incorporates textual features, visual tampering features and cross-modal similarity in fake news detection.
- **MCAN**(Wu et al. (2021)), which stacks multiple co-attention layers to fuse the multimodal features.

- **CAFE**(Chen et al. (2022)), which measures cross-modal ambiguity to help adaptively aggregate unimodal features and cross-modal correlations.

- **LIIMR**(Singhal et al. (2022)), which leverages intra and inter modality relationships for fake news detection.

- **FND-CLIP**(Zhou et al. (2023)), which uses two pre-trained CLIP encoders to extract the deep representations from the image and text.

- **CMC**(Wei et al. (2022)), which transfers cross-modal correlation by a novel distillation method.

4.2. Overall Performance

Table 1 showcases the performance comparison of MMFND and other methods on the Weibo and Twitter datasets. The best performances are indicated in bold, and the second best in underlined text. On each dataset, MMFND significantly outperforms all compared methods in both Acc and F1-score, demonstrating the effectiveness of our proposed model. Specifically, MMFND achieves an overall accuracy of 93.5% on the Weibo dataset, which is an increase of 2.7%, showcasing its advantage in handling Chinese fake news detection tasks. Similarly, on the Twitter dataset, MMFND also displays strong performance, with an accuracy of 89.6%, an increase of 6.5%, significantly higher than most other methods, indicating that MMFND can effectively handle cross-language and multi-domain news content.

Compared to CMC (Wei et al. (2022)), which performs second best overall, CMC demonstrates significantly lower accuracy on real news compared to its performance on fake news. This is attributed to the complexity of the CMC model, which leads to overfitting on fake news samples. In contrast, our method excels in detecting both real and fake news due to the incorporation of a multi-task network for feature sharing, thereby enhancing the model's generalizability and robustness. Many methods in fake news detection, such as EANN (Wang et al. (2018)) and MVAE (Khattar et al. (2019)), rely solely on obtaining fused features through direct concatenation or attention mechanisms. Although widely used, these methods might lack the discriminative power needed to effectively distinguish between true and false news if the independently extracted text and visual features are not in the same semantic space. CAFE (Chen et al. (2022)) employs a cross-modal alignment approach to train encoder models that can map text and visual data into a shared semantic space. The fused features obtained from aligned text and image inputs are subsequently used for classification purposes. However, due to a limited number of available datasets and suboptimal labeling methods used during training, the effectiveness of the encoding process may be hindered. This still results in significant semantic gaps between text and image features, which could impact overall classification performance. Our research differs from previous approaches by achieving optimal alignment of visual and linguistic representations through CLIP. Unlike direct alignment using contrastive learning methods, CLIP, by training on a massive corpus of image-text pairs, captures richer semantic information. Our findings indicate that the proposed model can capture highly refined cross-modal features.

Table 2: Ablation Studies.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	MMFND multimodal-only	0.851	0.845	0.888	0.866	0.858	0.805	0.831
	MMFND image-only	0.834	0.835	0.869	0.852	0.831	0.790	0.810
	MMFND text-only	0.863	0.869	0.888	0.879	0.854	0.830	0.842
	MMFND w/o C	0.884	0.886	0.897	0.880	0.852	0.866	0.884
	MMFND w/o F	0.914	0.910	0.938	0.924	0.920	0.885	0.902
	MMFND w/o A	0.921	0.920	**0.941**	0.930	0.923	0.897	0.910
	MMFND w/o S	0.906	0.899	0.914	0.907	0.914	0.898	0.906
	MMFND w/o P	0.870	0.868	0.902	0.885	0.874	0.832	0.852
	MMFND	**0.935**	**0.945**	0.934	**0.939**	**0.932**	**0.943**	**0.937**
Twitter	MMFND multimodal-only	0.839	0.839	0.858	0.848	0.840	0.818	0.829
	MMFND image-only	0.804	0.802	0.829	0.815	0.806	0.776	0.791
	MMFND text-only	0.822	0.816	0.850	0.833	0.829	0.793	0.811
	MMFND w/o C	0.846	0.845	0.865	0.855	0.847	0.826	0.836
	MMFND w/o F	0.875	0.872	0.892	0.882	0.878	0.856	0.867
	MMFND w/o A	0.882	0.878	0.899	0.888	0.885	0.863	0.874
	MMFND w/o S	0.862	0.859	0.866	0.863	0.864	0.857	0.861
	MMFND w/o P	0.838	0.835	0.860	0.847	0.842	0.814	0.828
	MMFND	**0.896**	**0.892**	**0.912**	**0.902**	**0.901**	**0.878**	**0.889**

4.3. Ablation Studies

To evaluate the effectiveness of each component of the proposed MMFND, we conducted ablation studies by removing certain parts from the entire model for comparison. The comparison variants of MMFND are implemented as follows.

- **MMFND multimodal-only**, only the features extracted from the cross-modal interaction task are used as final features.
- **MMFND text-only**, only the features extracted from the unimodal representations task of textual modality are used as the final features.
- **MMFND image-only**, only the features extracted from the unimodal representations task of visual modality are used as the final features.
- **MMFND w/o A**, removing the modality attention module and similarity weighting, directly concatenating the three features to obtain the final feature.
- **MMFND w/o F**, removing the fusion module, only the features extracted from the unimodal representations task of textual modality and visual modality are used as the final features.
- **MMFND w/o C**, removing all CLIP-related modules, only using BERT and ResNet to extract text and image features.
- **MMFND w/o S**, removing the similarity weighting component and fuse directly using the attention mechanism.
- **MMFND w/o P**, removing the PLE module, only concatenating CLIP-encoded multimodal features to form cross-modal features.

Table 2 shows the results of the ablation experiments, where the bold font indicates optimal and the underlined font indicates sub-optimal. From the results, (1) MMFND outperforms MMFND w/o P, indicating that PLE is effective in refining the features, the fused features obtained by this method have stronger expressive ability. (2) MMFND outperforms MMFND w/o S, indicating that weighting using similarity can effectively adjust the strength of cross-modal features. (3) MMFND outperforms MMFND w/o C, indicating that encoding the features using CLIP improves the model's ability to read and process the sample information. (4) MMFND outperforms MMFND w/o F, indicating that the cross-modal features obtained from feature refinement are effective and can make full use of the complementarity between different modalities. (5) MMFND outperforms MMFND w/o A, indicating that the use of similarity weighting and attentional weighting can effectively exploit the correlation between different modalities. (6) Comparing the performance of MMFND image-only, MMFND text-only and MMFND multimodal-only, it can be seen that the single visual information has fewer clues in fake news detection, while the textual information is richer, and the combination of visual features with the text can provide complementary information, which proves that the combined performance of multimodal features is better than that of the single modal features.

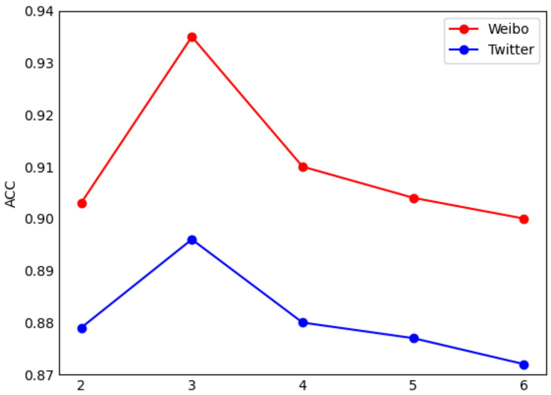


Figure 4: ACC with Different **E**.

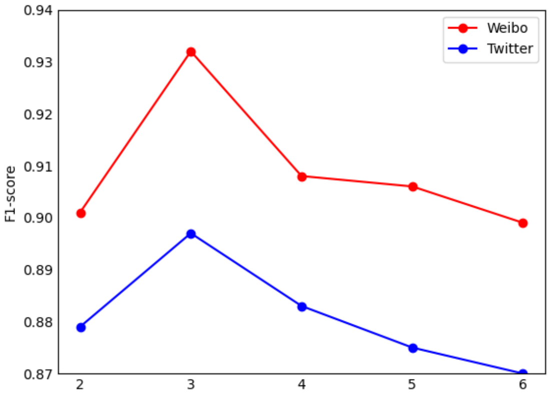


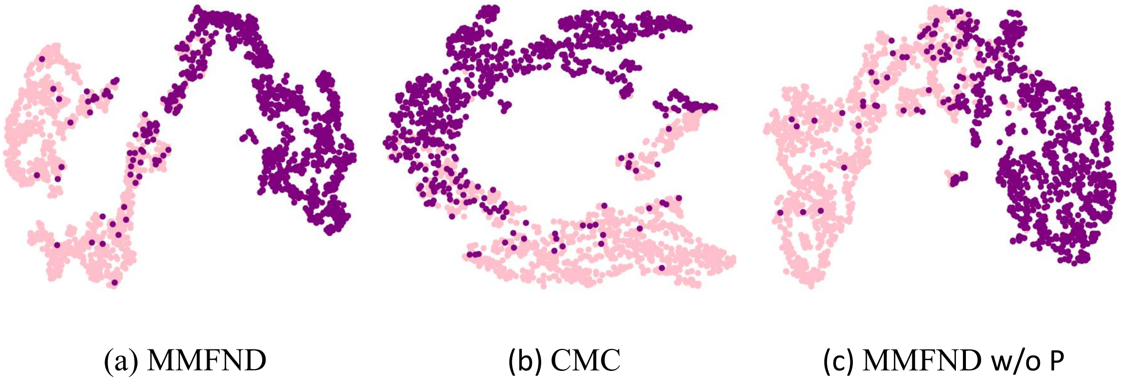
Figure 5: F1-score with Different **E**.

4.4. Hyperparameter Discussion

We enhance the model performance by adjusting the number of experts **E** in the PLE module. Figure 4 and Figure 5 shows the experimental results on the Weibo and Twitter datasets. The horizontal axis represents the number of experts and the vertical axis represents the ACC and F1-score for different numbers of experts, respectively. It can be clearly seen from the figure that the model performs best when the number of experts is 3. This may be because we use two unimodal features and one multimodal feature during feature aggregation. When the number of experts exceeds 3, the performance begins to decline. This indicates that more experts are not necessarily better, and the number of experts should be set appropriately according to different tasks.

4.5. T-SNE Visualizations

In Figure 6, we conduct an in-depth analysis of the proposed method using t-SNE (Van der Maaten and Hinton (2008)). The two colors in the figure represent the "true" and "false" labels, respectively. By comparison, it can be observed that in the MMFND model, the boundary between data points with different labels is significantly clearer than in CMC. This clear boundary indicates that the features extracted by our method are more discriminative compared to the existing best methods. Compared to MMFND w/o P, data with different labels in MMFND are significantly farther apart and have fewer outliers, suggesting that PLE-based refinement can improve the representation of fused features.


(a) MMFND

(b) CMC

(c) MMFND w/o P

Figure 6: T-SNE visualizations of the features before classifier that are learned by MMFND, CMC and MMFND w/o P on the test dataset of Weibo.

5. CONCLUSIONS

This paper introduces the Multi-Task Guided Multimodal Fusion for Fake News Detection framework (MMFND), which enhances cross-modal interaction and better utilizes cross-modal correlations in multimodal fake news detection. Our approach uses a multi-task network for feature refinement and paired CLIP encoders for modality alignment, measuring and adjusting the strength of cross-modal features to reduce ambiguity.

We also employ an improved SE-Net to adaptively weight and aggregate multimodal cues. Extensive experiments on two typical fake news datasets demonstrate that MMFND achieves high accuracy and F1-scores, validating its effectiveness. This framework offers a new perspective and method for multimodal fake news detection.

6. LIMITATIONS

MMFND faces two main limitations. First, we use pretrained models to extract features from news posts, which means the model's performance is largely dependent on the quality of the embeddings from these pretrained models. Although we fine-tune the ResNet networks during training, this approach does not adequately address the issue given the relatively small size of fake news datasets. Second, despite presenting a comprehensive theory and conducting extensive experiments for validation, the model's inference process remains a black box. Improving the model's interpretability is a motivation for our future work.

7. Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 6247074060), the Natural Science Foundation of Heilongjiang Province (No. LH2020F043), and the Fundamental Research Funds for Universities in Heilongjiang Province in 2024.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 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*, pages 14940–14949, 2022.

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

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.

Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

Juan Cao, Junbo Guo, Xirong Li, Zhiwei Jin, Han Guo, and Jintao Li. Automatic rumor detection on microblogs: A survey. *arXiv preprint arXiv:1807.03505*, 2018.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161, 2020.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14185–14193, 2021.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ole-Christoffer Granmo, Sondre Glimsdal, Lei Jiao, Morten Goodwin, Christian W Omlin, and Geir Thore Berge. The convolutional tsetlin machine. *arXiv preprint arXiv:1905.09688*, 2019.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2016.

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

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1867–1870, 2015.

Jiachen Ma, Jing Dai, Yong Liu, Meng Han, and Chunyu Ai. Contrastive learning for rumor detection via fitting beta mixture model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4160–4164, 2023a.

Jiachen Ma, Yong Liu, Meng Han, Chunqiang Hu, and Zhaojie Ju. Propagation structure fusion for rumor detection based on node-level contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

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*, volume 34, pages 13915–13916, 2020.

Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, pages 726–734, 2022.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 269–278, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Junda Wang, Jeffrey Zheng, Shaowen Yao, Rui Wang, and Hong Du. Tlfnd: A multimodal fusion model based on three-level feature matching distance for fake news detection. *Entropy*, 25(11):1533, 2023.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng 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)*, pages 4733–4737. IEEE, 2022.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610, 2021.

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

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

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, pages 2413–2419, 2022.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer, 2020.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE, 2023.

Peican Zhu, Jiaheng Hua, Keke Tang, Jiwei Tian, Jiwei Xu, and Xiaodong Cui. Multimodal fake news detection through intra-modality feature aggregation and inter-modality semantic fusion. *Complex & Intelligent Systems*, pages 1–13, 2024.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *Acm Computing Surveys (Csur)*, 51(2):1–36, 2018.