Cross-Modal Fine-Grained Interaction Fusion in Fake News Detection

Zhanbin Che, GuangBo Cui*
College of Computer, Zhongyuan University of Technology,
Zhengzhou, Henan 450007, China

***Abstract*—The popularity of social media has significantly increased the speed and scope of news dissemination, making the emergence and spread of fake news easier. Current fake news detection methods often ignore the correlation between text and images, leading to insufficient modal interaction and fusion. To address these issues, a cross-modal fine-grained interaction and fusion model for fake news detection is proposed. Specifically, this study addresses the correlation problem between text and image modalities by designing an interaction similarity domain. It extracts features of text word weight distribution using an attention mechanism network, guides the features of different regions of the image, and calculates the local similarity between the two. This approach analyzes positive and negative correlations between modalities at a fine-grained level, thereby strengthening the intermodal connection. Additionally, to tackle the problem of insufficient fusion of semantic feature vectors between text and images, this paper designs a fusion network that employs improved encoding and decoding using a Transformer for inter-modal information fusion, achieving the final multimodal feature representation. Experimental results show that our proposed method achieves excellent performance on WeiboA and Twitter, with accuracies of 88.2% and 89%, respectively, outperforming the benchmark model in several evaluation metrics.**

***Keywords*—Fake news detection; attention mechanism; multimodal feature fusion; local similarity**

I. INTRODUCTION

With the advent of the Internet, a multitude of online social media platforms such as Twitter, Weibo, Shake, and Shutterbug have experienced unprecedented growth [1]. These platforms, characterized by their low operational costs, high efficiency, real-time capabilities, and the diverse nature of their content, have revolutionized traditional methods of information dissemination. Consequently, an increasing number of individuals are gravitating towards these platforms for information acquisition and personal life sharing, thus diversifying the modalities of information exchange. However, this evolution has inadvertently facilitated the genesis and proliferation of fake news. Online fake news not only seriously impacts the audience and weakens the authority and credibility of mainstream media institutions, but also brings risks in many aspects, including economic and political [2]. A pertinent example of the detrimental impact of misinformation is the wide dissemination of spurious content during the U.S. Capitol riots in January 2021, which obscured the factual narrative and intensified societal polarization. Hence, it is imperative to devise and implement sophisticated methods for the detection and containment of fake news to mitigate its adverse effects on public discourse and social harmony.

In the realm of fake news detection, traditional approaches have predominantly centered around the verification processes conducted by domain experts or credible institutions [3]. While this strategy is commendably precise, its feasibility has been compromised by the contemporary influx of voluminous information and the escalation of operational costs. In response to these challenges, academia has ventured into the realm of manual feature extraction, focusing on lexical, syntactic (e.g., structure and grammar), and semantic (encompassing rhetorical techniques, thematic consistency, and emotive expressions) aspects. These extracted features are then amalgamated with established machine learning models like decision trees and support vector machines to discern deceptive information [4][5][6]. Nevertheless, this manual feature extraction method often falls short in grasping intricate semantics and complex narratives, thus limiting the overall performance of detection systems. Given the potentially severe repercussions of misinformation spread, the academic community is actively engaged in advancing the capabilities and accuracy of these detection mechanisms. Consequently, refining methodologies for the accurate detection of fake news has emerged as a focal area of research, drawing significant scholarly interest and resource investment.

The evolution of deep learning has demonstrated substantial efficacy across diverse sectors, marked by its capacity for autonomous feature detection, advanced representational learning, and extensive generalization abilities. In the context of the multifaceted nature of news content, research initiatives are increasingly focused on deriving complex intermodal representations through deep neural networks. Yet, a predominant share of current methodologies relies on leveraging pre-trained models for feature extraction, followed by a simplistic concatenation to amalgamate multimodal features, often overlooking the critical nuances in informational content across different modalities. Such unimodal feature extraction methodologies inadequately harness the comprehensive information available from varied modalities, consequently impeding the effective formation of intermodal linkages [7][8][9]. Furthermore, In complex scenarios where the image and text do not match, for example, such as the example depicted in Fig. 1, presenting a fake news report about a new fish product, where most regions of the image depict fish characteristics, while a small portion does not. The primary regions of the image align with the text, whereas the secondary regions do not. Relying solely on the overall similarity between

*Corresponding Author

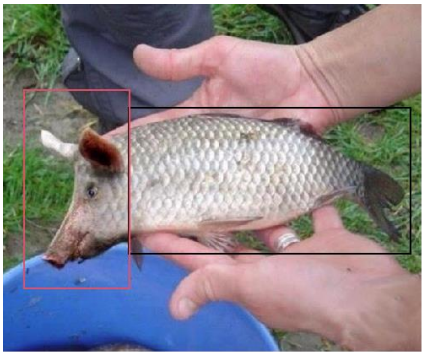
the text and the image for calculations may lead to erroneous model judgments and impact its performance.

To address the aforementioned issues, this paper proposes a cross-modal fine-grained interactive fusion model for false news detection. To tackle the problem of insufficient interaction between modalities, the model employs an attention mechanism network to extract text word weight distribution features, which guide the extraction of features from different regions of the image, thereby strengthening inter-modal connections. In complex scenarios of graphical inconsistency, this study utilizes text word weight distribution features and image region features for similarity calculation, obtaining local similarity features that enable a more granular analysis of positive and negative correlations between modalities. This increases the likelihood of the model accurately extracting relevant features. Additionally, to overcome the challenge of directly merging text and image semantic feature vectors in modality fusion, a fusion network is designed to effectively integrate modal information, resulting in a comprehensive multimodal feature representation. By refining these modalities, the detection performance of the model is significantly improved. The main contributions of this paper are as follows:

1) An interactive similarity domain is designed to extract text word weight distribution features using a network of attention mechanisms to guide different levels of image feature extraction and to obtain fine-grained local similarity features between modalities, aiming to strengthen inter-modal connections and enhance the effectiveness of modal features.

2) A novel fusion network, featuring an improved Transformer dual-encoder architecture, has been devised to meticulously extract deep semantic cues from multimodal fake news content. This architecture facilitates the realization of a highly accurate multimodal feature representation, optimizing the detection and analysis of counterfeit information across varied modalities.

3) Through extensive comparison and ablation experiments with benchmark models such as the classical EANN, conducted on two popular multimodal fake news detection benchmark datasets, WeiboA and Twitter, the CFIF model demonstrates superior performance across most evaluation metrics.



Text: New species of fish found at Arkansas

Fig. 1. Some examples of multimodal fake news.

The rest of this paper is as follows, Section II review previous studies. Section III discusses the methodology. Section IV presents experimental setup. Section V describes the results of the experiment and discusses. Finally, conclusion presents in Section VI.

II. Related Works

Fake news detection employs news article content, social context, and external knowledge to assess news authenticity. This section introduces two primary approaches from the perspective of modality quantity: unimodal and multimodal fake news detection. In terms of effectiveness, multimodal detection demonstrates superior performance due to its richer and more comprehensive information. However, simple modality fusion and insufficient modality interaction cannot satisfy the current research needs, as the model requires features with finer granularity and greater generalizability.

A. Unimodel-based Fake News Detection

In history, news predominantly existed in textual form, encapsulating the author's perspectives, emotions, and stylistic choices. Leveraging this information, lexical, syntactic, and semantic features can be extracted. Therefore, the core of unimodal machine learning detection techniques lies in adeptly constructing and filtering features to accurately represent textual news information. Horne et al. [4] categorized text features into three main groups—style, complexity, and psychological traits—analyzing them at the word level with a Support Vector Machine (SVM) model to identify fake news. Similarly, Perez-Rosa et al. [5] manually compiled a set of text features at the word level, comprising n-grams, punctuation, psycholinguistic attributes, and generative rules, and employed the SVM model for fake news detection. However, this method faces challenges in feature interpretability and handling the variability and diversity of fake news. Castillo et al. [6] devised a suite of linguistic features, including question marks, emoticons, emotional words, and pronouns, to evaluate the credibility of tweets and detect fake news. Although manual feature extraction progress in detecting fake news, the required targets and features differ among news types, distribution channels, and dissemination routes. Consequently, extracting unique features for each news type proves to be both resource-intensive and time-consuming. Moreover, to preserve the stability and accuracy of detection outcomes, feature extraction techniques must be regularly updated and refined to accommodate evolving news events, leading to an inevitable rise in costs.

Deep learning technology has demonstrated its robustness and effectiveness across various domains. Its primary strengths lie in its capacity to autonomously extract data features, superior representation learning, and broad generalization capabilities. Ma et al. [10] explored how deep neural networks, utilizing Word-Embedding and RNN models, could represent news to enhance detection efficiency and accuracy, thus providing innovative approaches for applying deep learning in journalism. Volkova et al. [11] analyzed tweet texts using linguistic markers, social graphs, bias, subjectivity, and ethical features, employing CNN and LSTM networks to categorize information, yet this method did not enhance model performance, even with the integration of grammatical and

syntactic elements. Chawda et al. [12] highlighted the significance of context in text categorization by employing a Recurrent Convolutional Network (RCNN) with an LSTM network, achieving improved accuracy. Hansen He et al. [13] proposed a fake news detection model based on feature aggregation, employing a BiLSTM network to extract global temporal features and a CNN network for word or phrase features within a window, thus enhancing the model's generalization capability.

With the evolution of the Internet and the diversification of news forms on social media, some scholars have shifted their focus to image analysis for detecting fake news. Qi et al. [14] , developed a CNN-based network to identify complex patterns in fake news images within the frequency domain and a multi-branch CNN-RNN model to extract visual features across various semantic levels in the pixel domain. They integrated these features from both domains using an attention mechanism to enhance detection. However, this method heavily depends on sophisticated visual feature extraction, posing challenges in identifying subtle alterations used by fake news creators, potentially compromising detection accuracy. Xue et al. [15] introduced the MVFNN model, comprising a visual modality module, a visual feature fusion module, a physical feature module, and an integration module, all working synergistically for fake news image detection. Zhou et al. [16] proposed a method to identify tampered regions using a dual-stream Faster R-CNN network: one stream processes RGB images to extract features like contrast differences, while the other analyzes noise inconsistencies from the model's filter layer, with both feature sets subsequently fused for detection.

Unimodal methods have advanced significantly in detecting fake news but exhibit several limitations. Primarily, they rely on a single modality, such as text or image, neglecting multi-source information, which compromises detection accuracy due to the multi-modal nature of fake news. Furthermore, these methods are susceptible to adversarial attacks, as attackers can bypass detection by crafting sophisticated false information. Additionally, unimodal methods often overlook inter-modal correlations, resulting in incomplete information capture. Lastly, they struggle with cross-modal fake news, where fake news spreads across different modalities like social media, news articles, and images.

B. Multimodal-based Fake News Detection

In response to the diversity of news and the limitations of unimodal false news detection, researchers have shifted towards multimodal approaches. Initially, these methods separately extracted unimodal features, combining them sequentially, as illustrated in Fig. 2. Jin et al. [7] were the pioneers in proposing a multimodal fake news detection framework, utilizing the LSTM model for text and the VGG-19 model for image feature extraction, followed by sequential integration for classification. Chen et al. [8] implemented DeepFM-a blend of deep learning and factorization machines-to assess social news features, Text-CNN, and VGG-19 for textual and visual feature extraction, merging these elements to derive multimodal features for classification. Wang et al. [9] also employed Text-CNN and VGG-19 to process text and image data, respectively, but enhanced the approach by adding an event discrimination module and applying Adversarial Neural Networks, which significantly improved detection efficacy.

While the aforementioned methods have notably enhanced the performance of fake news detection compared to unimodal approaches, they have not fully leveraged the complementary information across modalities. To address this, researchers have developed advanced methods. Zhou et al. [17] proposed a similarity-aware model that identifies discrepancies between text and images in fake news, employing Text-CNN for text feature extraction and an image2sentence model to transform image features, with classification subsequently based on the similarity between these elements. Song et al. [18] employed a combination of multiple attention mechanisms and the pre-trained VGG-19 model to selectively cross-learn information from different modalities using a bidirectional cross-attention mechanism, preserving the original feature information. This approach has proven effective across four datasets.

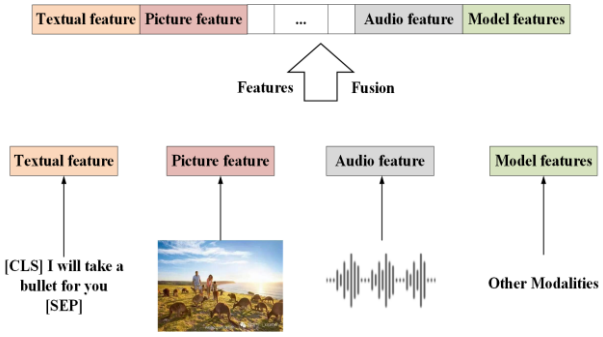


Fig. 2. Feature fusion diagram.

In conclusion, recent research in fake news detection has increasingly leveraged both image and text information, achieving notable success in identification. Nonetheless, these methods continue to confront various challenges：

1) Insufficient interaction between modalities: Most multimodal fake news detection models extract high-level features from images by designing specific models for different modal data, e.g., using pre-trained VGG models, but the lack of effective interactions before fusing these modalities restricts the ability of the model to fully utilize the information between modalities, which in turn affects the performance.

2) Insufficient fine-grained analysis: Most models use cosine similarity for intermodal similarity calculations, which may lead to the selection of incorrect features and the failure to capture data details, thereby affecting the model's accuracy and reliability.

3) Feature fusion methods are comparatively simple: Classic fake news detection models like EANN adopt a straightforward concatenation strategy, which not only increases computational complexity but also results in information redundancy. Outer product fusion may lead to excessively large dimensions of output features, thereby raising the risk of dimensional explosion.

To tackle these challenges, this study develops a multimodal fake news detection model emphasizing the

detailed analysis of modalities and their interactive fusion to improve semantic feature extraction, thereby enhancing the accuracy of fake news detection.

III. METHODOLOGY

A. Overview of the Model

In this study, we present a false news detection model cross-modal fine-grained interaction fusion. Addressing the challenges of inter-modal interaction and fine-grained analysis, the model employs a text attention mechanism to guide the generation of image features. Given BERT's [19] strong feature extraction capabilities, which may lead to local optimization of text features, therefore, the Text-CNN [20] model's sparsity is leveraged to filter noise and capture text features at various granularities. Moreover, recognizing the varying information and importance across image regions, we introduce a weighted region division method using image segmentation technology, followed by ResNet-50 for detailed feature extraction from the image. To resolve the issue of simplistic feature fusion, the model incorporates a fusion network that integrates text and image features, further enhanced with local similarity metrics to produce the final fused features. The model comprises three core components: the feature extractor for text and images, the feature fusion mechanism, and the feature discriminator, with its comprehensive framework depicted in Fig. 3.

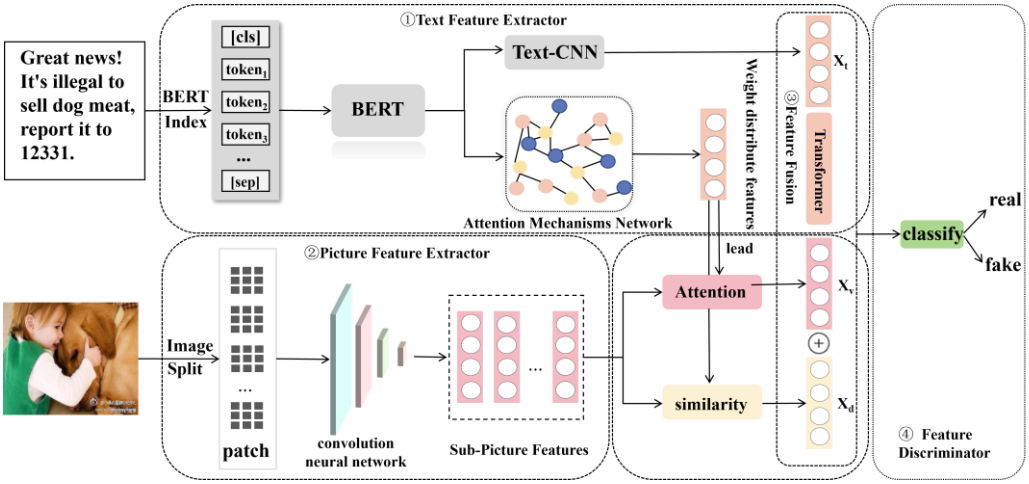


Fig. 3. CFIF model structural framework.

B. Multimodal Feature Extraction

1) Textual Feature Extraction: Text feature extraction is crucial for detecting fake news. In this study, we employed the pre-trained ROBERT [21] model for text labeling and initial feature extraction. ROBERT, an advanced variant of BERT, dynamically generates vector representations of words in various contexts, addressing the context-independence issue inherent in Word2vec [22]. Moreover, ROBERT utilizes a larger corpus and undergoes more extensive training than BERT. Additionally, it implements a dynamic masking strategy that generates a new mask pattern each time a sequence is processed, enabling the model to gradually adapt to various masking strategies and learn diverse linguistic representations. This adaptability across different domains makes ROBERT particularly suitable for our research needs.

Initially, the text $T=\{W_1,W_2,W_3,...,W_n\}$, is represented, where W_i denotes the i th word in T. $T'=\{[CLS],Token_{W_1},Token_{W_2},...,Token_{W_n},[SEP]\}$ is the result of the BertTokenizer segmenting the text into tokens. Subsequently, these tokens are converted into their respective IDs and fed into the ROBERT model to generate the word vectors $V=\{V_{[CLS]},V_{W_1},V_{W_2},...,V_{W_n},V_{[SEP]}\}$,where V_i represents the vector for the ith word. These vectors are then input into the Text-CNN model, which employs various convolutional kernels and sliding windows to further extract the semantic information X_t from the text, as delineated in Eqs (1)-(3).

$$T'=BertTokenizer(T) \tag{1}$$

$$V=ROBERT(T') \tag{2}$$

$$X_t=Text-CNN(V) \tag{3}$$

Besides the text being the core element of the news event, the image is also a significant modality. Therefore, it is necessary not only to input the word vector V into the $Text-CNN$ to obtain a comprehensive textual representation but also to feed V and the mask into the Attention Mechanism Network to learn the distribution of word weights in the text. The aim is to update the weight distribution of the original image and adjust the weight of the semantic information in the image, as shown in Eq (4).

$$X_c=Mask\_Attention(V) \tag{4}$$

2) Visual Feature Extraction: Given that images are inherently more intuitive than text, making them easier to understand and recall [23][24] , their widespread adoption in news articles has become customary. This underscores the criticality of efficiently extracting image data in the detection of fake news. Convolutional networks, particularly VGG [25] and ResNet models, have emerged as efficient tools for

extracting these crucial image features. While the VGG model excels in extracting image features with greater accuracy, it demands higher computational resources due to its larger memory footprint and parameter count. Moreover, the VGG model is plagued by the issue of gradient vanishing. Thus, for this study, we opted for the ResNet-50 model from the ResNet family. Not only does it achieve remarkable progress in accuracy and minimize loss, but it also resolves the gradient vanishing problem, boasts a deeper network structure, and is particularly well-suited for classification tasks.

Initially, establish a data conversion pipeline to standardize the image dimensions to (224×224) and convert them to RGB three-channel style. Let C represent the original image. Then, divide the image into multiple identical regions $\{P_1, P_2, P_3, ..., P_N\}$, where each P_i represents a portion of the image C with dimensions of (32×32), resulting in a total of 49 copies. Next, each P_i undergoes feature extraction using ResNet-50, yielding subgraph features denoted as Sub_P_i. Additionally, adjust the information of each image copy using the textual weight distribution information X_c acquired in Section Ⅲ. *B. 1).* Finally, these adjusted features are weighted, summed, and consolidated to obtain the refined image information X_v, as depicted in Eqs (5)-(8):

$$Split(C) = P_1, P_2, P_3, ..., P_N \tag{5}$$

$$Sub\_P_i = \sigma(W \times Resnet(P_i) \pm b) \tag{6}$$

$$a_i = SoftMax(S(X_c, Sub\_P_i; \Phi)) \tag{7}$$

$$X_v = \frac{\sum_{i=1}^{N} a_i \cdot Sub\_P_i}{N} \tag{8}$$

where $S(\ ,\ ;\ \Phi\ )$ represents a mapping network, a_i signifies the weight vector determining the significance of the subgraphs, N denotes the quantity of subgraphs, while W and b denote the parameters of the fully connected layer, and $\sigma(\cdot)$ denotes the activation function.

C. Local Similarity Feature Extractor

Fake news detection usually involves some complex cases where the image and text do not match. For example, if the body of an image matches the text semantically, while other parts do not, directly calculating their similarity may lead to detection errors. To cope with such problems, in this paper, we use cosine similarity to calculate the text weight distribution feature X_c and the image region feature vector Sub_P_i. The similarity of the text weight distribution feature X_c and the image region feature vector Sub_P_i, after that, the normalization process is performed to obtain the similarity contribution of each part, and finally the weighted sum is obtained to obtain the local similarity feature. As shown in Eq. (9)-(11):

$$part\_similarity\_i(X_c, Sub\_P_i) = \frac{X_c \cdot Sub\_P_i}{\max(\| X_c \|_2 \cdot \| Sub\_P_i \|_2, \varepsilon)} \tag{9}$$

$$w_i = \frac{\exp(part\_similarity\_i)}{\sum_{i=1}^{N} \exp(part\_similarity\_i)} \tag{10}$$

$$X_d = \sum_{i=1}^{N} w_i \cdot part\_similarity\_i \tag{11}$$

where X_c is the word weight feature obtained in Section Ⅲ. *B. 1).*, $\|\cdot\|_2$ is the l_2 normalization, w_i is the proportion of subgraphs， X_d is a local similarity feature.

D. Multimodal Feature Fusion

Using text features X_t, image features X_v and local similarity features X_d, designing an efficient feature fusion method so as to obtain effective multimodal features is the key to realize fake news detection. If X_t, X_v and X_d are simply spliced together may lead to information redundancy as well as dimension explosion, and the outer product fusion method may lead to multimodal information asymmetry and high computational complexity.

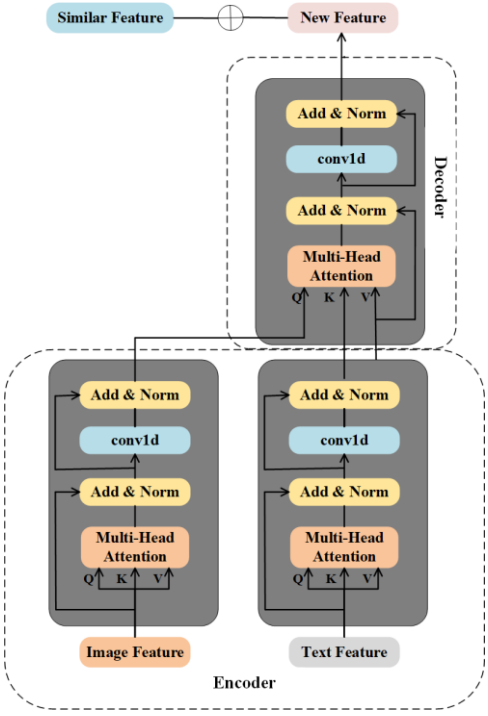


Fig. 4. Multimodal fusion framework diagram.

To avoid the above problems, the fusion network is designed in this study, firstly, the Encoder framework automatically focuses on the key features of text and images through the multi-head self-attention mechanism. At the same time, the residual network is introduced to preserve the original features, and the feed-forward network is replaced with a one-dimensional convolutional network in order to reduce the computational complexity. Second, the Decoder framework adopts the same structure as Encoder, but the difference is that its input combines text (primary) and image (secondary) features and fuses them with local similarity features to obtain

a multimodal information representation. The feature fusion apparatus, as depicted in Fig. 4 below.

1) Encoder: Since picture modality and text modality coding are consistent, take the text modality coding process as an example, firstly, the text modality feature vector X_t is taken as the input, and the self-attention mechanism is carried out to continuously strengthen the information of the text modality. the Query vector $Q_t = X_t W_Q$,the Key vector $K_t = X_t W_K$,and the Value vector $V_t = X_t W_V$, where $W_Q \in R^{d_t \times d_k}$, $W_K \in R^{d_t \times d_k}$, $W_V \in R^{d_t \times d_v}$, d_t is the sequence length of the text modality, d_k is the dimension of the Query vector and Key vector, and d_v is the dimension of the Value vector. In this paper, we use Multi-Head Attention (MHA) mechanism to capture different attention information within and between modalities, and utilize the subspace of the multi-head matrix to express modality information from different perspectives. Multihead Attention is multiple independent Attention computations that are then stitched together. The computational process of the Multihead Attention mechanism is shown in Eq. (12)-(13):

$$head_i = Attention(Q_t W_i^Q, K_t W_i^K, V_t W_i^V) \tag{12}$$

$$MHA(Q_t, K_t, V_t) = Concat(head_1, ..., head_h) W^O \tag{13}$$

where $W_i^Q \in R^{d_t \times d_k}$, $W_i^K \in R^{d_t \times d_k}$, $W_i^V \in R^{d_t \times d_v}$, $W^O \in R^{hd_v \times d_t}$, $d_k = d_v = d_t / h$.

The feature vector X_t of the text modality undergoes the Multi-Head Attention (MHA) mechanism. The new vector representation X_t' is obtained through residual connection and layer normalization, and the computational process is depicted in Eq. (14) as follows.

$$X_t' = LayerNorm(X_t + MHA(Q_t, K_t, V_t)) \tag{14}$$

Next, X_t' is the input to the second sublayer, which consists of a one-dimensional convolutional network, residual connections, and layer normalization operations. Subsequently, the output of the Encoder can be obtained. The computational process is then depicted in Eq. (15):

$$\overline{X}_t = LayerNorm(X_t' + Conv1d(X_t')) \tag{15}$$

where $\overline{X}_t \in R^{d_t \times d_{model}}$, d_t is the sequence length of the text modal and d_{model} is the size of the Embedding.

The same operation is performed to obtain the image modal features $\overline{X}_v$.

2) Decoder: The primary function of the Decoder is to facilitate cross-modal interaction between the text modality features enhanced by the Encoder and the picture modality features in order to obtain effective multimodal features. Specifically, the operation involves inputting the text modality feature $\overline{X}_t$ outputted from the Encoder as Key and Value vectors, and the picture modality feature $\overline{X}_v$ outputted from the Encoder, as the Query vector into the Decoder, collectively obtaining the mixed text and picture features MTF . The specific formula is illustrated in Eq. (16)-(19):

$$Q_v = \overline{X}_v W_Q \tag{16}$$

$$K_t = \overline{X}_t W_K \tag{17}$$

$$V_t = \overline{X}_t W_V \tag{18}$$

$$MTF = \text{Softmax}(\frac{Q_v K_t^T}{\sqrt{d_k}}) V_t \tag{19}$$

where $W_Q \in R^{d_v \times d_k}$, $W_K \in R^{d_t \times d_k}$, $W_V \in R^{d_t \times d_v}$, d_v is the sequence lengths of the visual modality, d_t is the sequence length of the textual modality, d_k is the dimension of the Query and Key vectors.

Subsequently, the textual modality feature vector $\overline{X}_t$ and the mixture feature vector MTF are processed through residual connection and layer normalization to obtain the vector MTF' as the output of the first sublayer. The computational formula is illustrated in Eq. (20).

$$MTF' = LayerNorm(\overline{X}_t + MTF) \tag{20}$$

Next, MTF' is inputted into the second sublayer to obtain the output $\overline{MTF}$ of the Decoder. Afterward, it is combined with X_d to yield the final modal features. The specific equations are depicted in Eq. (21)-(22):

$$\overline{MTF} = LayerNorm(MTF' + \text{Conv1d}(MTF')) \tag{21}$$

$$HMF = \overline{MTF} \oplus X_d \tag{22}$$

where Conv1d is the one-dimensional convolutional layer, X_d is the local similarity feature, and HMF is the final multimodal feature.

E. Fake News Detector

In this study, we frame the fake news detection task as a binary classification problem. We input the HMF (multimodal fusion feature) from Section Ⅲ. *D. 2)* into a neural network consisting of a single hidden layer. The dimension of this hidden layer is configured to be twice that of the multimodal features. Subsequently, a Softmax function is employed to compute the probability distribution, with the category exhibiting the highest probability serving as the ultimate classification outcome, as illustrated in Eq. (23):

$$P = \text{Softmax}(W * HMF + b) \tag{23}$$

where P represents the probability that the prediction is false, while W and b denote the parameters of the fully connected layer. The real news is labeled as 0, and the fake news is labeled as 1.

The cross-entropy loss function typically exhibits better gradient properties during optimization, allowing for more effective parameter updates during back-propagation. The gradient computation of the cross-entropy loss function involves comparisons between different classes, which guides parameter updates more effectively, leading to faster convergence to the optimal solution during training. In contrast, the gradient computation of the mean squared error function is more influenced by the errors between predicted and true values, which may sometimes lead to gradient vanishing or exploding, resulting in unstable parameter updates and affecting the training effectiveness of the model. Therefore we choose cross entropy as the loss function for this study. Hence we opt for cross entropy as the loss function in this study, as illustrated in Eq. (24).

$$H(y, p) = -\sum[y \log p + (1-y)\log(1-p)] \quad (24)$$

Where $H(y, p)$ represents the binary cross-entropy loss, y stands for the true category label, p signifies the predicted probability of the model, denoting the probability that the sample belongs to category 1, and log denotes the natural logarithm.

IV. Experimental Setup

A. Datasets and Evaluation Metrics

1) Datasets: To validate this study, two datasets, WeiboA and Twitter, were selected for experimentation. These datasets, representing both languages, were used to demonstrate the generalization ability of the model. Here:

WeiboA Dataset: Compiled by Jin et al. [26], this dataset captures all verified false news from May 2012 to January 2016 through the official microblogging rumor debunking system. Primarily consisting of articles reported by ordinary users, they undergo verification by a forensic group comprised of reputable users to determine their veracity. For authentic news text, articles verified by Xinhua News Agency, China's authoritative news agency, are utilized. The study aims to discern multimodal information; therefore, text-only posts and duplicate or low-quality images are excluded.

The Twitter dataset [27] is employed for false news discrimination and comprises a development set and a test set. Each data point in the dataset includes textual content, visual content (image/video), and relevant social context. As this study focuses on the visual modality of images, samples containing only video data are excluded. The development set is utilized as the training set, while the test set serves as the evaluation set.

The length of the text needs to be processed under the premise of ensuring that the real data is balanced with the false data. Inconsistent text length will affect the performance of the model, so longer or shorter data need to be eliminated, and finally, the data set is statistically analyzed to obtain the statistical information of the data as shown in Table I. In addition, Fig. 5 and Fig. 6 give examples of images and corresponding texts in the datasets.

TABLE I. Distribution of Each Dataset

Datasets	Originating data	Contains image data	Final data
WeiboA	9528	7723	7713
Twitter	13136	13136	13136

(A) Fake News (B) Real News

Fig. 5. Examples of real and fake news in the twitter dataset.

(a) Fake news (b) Real news

Fig. 6. Examples of real and fake news in the WeiboA Dataset.

2) Evaluation Metrics: The fake news detection models in this study fall under the classification model category. Evaluation metrics commonly employed to gauge model performance are presented in Eq. (25):

$$F_\beta = (1+\beta) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (25)$$

where, *Precision* measures the proportion of correctly identified positive cases, while *Recall* measures the proportion of actual positive cases correctly identified. Additionally, the value of β plays a crucial role in balancing *Precision* and *Recall*. Specifically:

(1) When $\beta = 1$, the F_β metric equates to the F_1 metric, signifying equal importance of *Recall* and *Precision*.

(2) When $\beta > 1$, *Recall* holds more significance than *Precision*.

(3) When $\beta < 1$, *Precision* has a greater impact than *Recall*.

In this study, β is set to 1, indicating equal importance of *Recall* and *Precision*. The equation at this point is shown in (26):

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{26}$$

B. Experimental Details

The experiments are conducted in a Python 3.8 programming environment using the PyTorch deep learning framework to build and train the fake news detection model. To prevent overfitting and enhance model robustness, Dropout is applied in the fully-connected layer. Additionally, the EarlyStop strategy is employed during training, and Adam is utilized as the optimization function. The specific parameters are detailed in Table II.

TABLE II. DETAILS OF EXPERIMENTAL PARAMETERS

Parameters	Value
Epoch	50
Learing Rate	0.0005
Dropout	0.5
Batch Size	32
Optimizer	Adam
Window Size	[2,3,4,5]
Hidden Layer	64
Number of heads	4

C. Baseline

In order to highlight the performance of the model, this study will compare the parameters (accuracy, precision, recall, and F1 value) with the current effective model (benchmark model), and the following benchmark models are involved in this study:

- **Textual**. Utilizes various convolutional kernels to extract text features and performs simple concatenation for classification.
- **Visual**. Utilizes a pre-trained ResNet50 model to extract solely image features for classification.
- **EANN [9]**. A classic multimodal fake news detection model. It utilizes a Text-CNN model to extract text features and a pre-trained VGG-19 model to extract image features. These features are concatenated and fed into the fake news detection model for classification. Additionally, it incorporates Adversarial Neural Networks for event discrimination.
- **SAFE [17]**. Utilizes a Text-CNN model to extract features from text and images. It employs an image2sentence model for modal transformation of images before extracting them with the Text-CNN model. Finally, it calculates the similarity between text and images for classification.
- **MVAE [28]**. Utilizes Bi-LSTM and VGG-19 for text and image feature extraction respectively. The features are then concatenated to form multimodal information, and a variational autoencoder (VAE) efficiently captures the complex structure and relationships of multimodal data for classification.
- **CARMN [18]**. Employs multi-head attention and pre-trained VGG-19 to learn news text and image features. Based on a bidirectional cross-attention mechanism, it selectively learns information from one modality to another and combines the residuals to preserve the original feature information.
- **DCNN [8]**. Integrates DeepFM with the FM algorithm to learn news social features. Text-CNN and VGG19 are employed to learn news text and image features, which are then concatenated to obtain multimodal features.
- **MCNN [29]**. Utilizes deep learning combined with the ELA algorithm to extract image tampering features. BERT and Bi-GRU extract textual feature sequences, while ResNet50 and an Attention mechanism extract visual semantic features. These features are used to explore the consistency of multimodal content after extraction.
- **Roberta+CNN [30]**. This framework integrates a specialized convolutional neural network model for image examination and a sentence transformer for textual evaluation. Characteristics derived from visual and textual sources are merged via dense layers, ultimately converging to forecast deceitful visuals.
- **BDANN [31]**. Textual characteristics in BDANN are derived from a pre-trained BERT model, whereas visual traits are acquired through a pre-trained VGG-19 model. Reliance on particular events is lessened by integrating a domain classifier.

V. RESULTS AND DISCUSSION

A. Experimental Results

Table III presents the results of both the benchmark model proposed in the previous subsection and the model proposed in this study, using equivalent evaluation metrics, on the WeiboA and Twitter datasets.

Upon analyzing the experimental results from both datasets, several key conclusions emerge. Firstly, across both the WeiboA and Twitter datasets, this study's model outperforms alternative methods. This superiority suggests the effectiveness of leveraging word weight features from news text to enhance the semantic information of images, alongside employing a fine-grained approach for extracting local similarity features. Moreover, the multimodal features extracted in this study exhibit greater effectiveness compared to alternative methods.

In the WeiboA dataset, the performance of the Text-CNN model surpasses that of the Visual-CNN model. Remarkably, even when considering unimodal information, the Text-CNN model's performance closely rivals that of the EANN model utilizing multimodal information. This underscores the pivotal role of textual information in journalism, given its rich

semantic, emotional, and contextual content. The Visual-CNN model's underperformance may be attributed to subpar image quality or insufficient key information, leading to noise in the extracted features.

In the multimodal scenario, the MVAE model exhibits the poorest performance, possibly attributed to its simplistic fusion of modal information lacking information redundancy resulting from modal interaction. Additionally, the performance decline could be due to the variable autocoder's sensitivity to hyperparameters such as variable dimensions and loss functions. Conversely, the inclusion of social scenario features in the DCNN model does not improve performance; rather, it decreases it. This decline may stem from the distant relation between social scenario features and text/picture features, introducing noise and consequently impacting model performance. Similarly, the EANN model, despite incorporating an event discrimination module to leverage adversarial learning for discarding event-specific features, faces performance challenges akin to the MVAE model due to its rudimentary fusion of model features. Moreover, completely discarding event-specific features risks losing vital event context. In contrast, the CARMN model outperforms the current model, leveraging an attention mechanism akin to this study. However, this study's comprehensive interaction of textual information with images leads to superior results compared to the CARMN model.

The model demonstrates superior performance on the Twitter dataset compared to the WeiboA dataset. This discrepancy is likely due to the smaller number of training samples available in the WeiboA dataset, resulting in insufficient information for effective learning. Consequently, the quality of multimodal features diminishes, resulting in marginally poorer performance on the Twitter dataset.

TABLE III. COMPARISON OF ACCURACY, PRECISION, RECALL, AND F1 FOR DIFFERENT BASELINES

Datasets	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
WeiboA	Textual	0.832	0.860	0.816	0.838	0.804	0.850	0.827
	Visual	0.668	0.686	0.688	0.687	0.648	0.645	0.646
	EANN	0.836	0.843	0.851	0.847	0.828	0.819	0.824
	MVAE	0.750	0.731	0.864	0.781	0.812	0.629	0.709
	CARMN	O.853	**0.891**	0.814	0.851	0.818	**0.894**	0.854
	DCNN	0.803	0.804	0.819	0.811	0.803	0.787	0.795
	Roberta+CNN	0.812	0.851	0.784	0.816	0.744	0.826	0.782
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	CFIF	**0.882**	0.883	**0.901**	**0.881**	**0.891**	0.873	**0.884**
Twitter	Textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Visual	0.596	0.695	0.518	0.593	0.524	0.700	0.599
	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
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	Roberta+CNN	0.853	0.821	**0.943**	0.877	0.913	0.745	0.820
	BDANN	0.830	0.810	0.630	0.710	0.830	**0.930**	**0.880**
	CFIF	**0.890**	**0.871**	0.940	**0.901**	**0.921**	0.833	0.872

B. Analysis of Ablation Experiments

To validate the efficacy of each module within the model proposed in this study, we conducted ablation experiments by disassembling each module of CFIF. These experiments aimed to explore the impact of each module on performance, focusing on the following variants:

1) CFIF-M: The modal interaction module is removed and the extracted word weight features are not involved in the generation of visual features. It is used to validate the effectiveness of the interaction module.

2) CFIF-L: Removes the locally similar features and the multimodal features include only the combination of textual features and visual features. Used to validate the effectiveness of the similarity module.

3) CFIF-F: Remove the Modified Transformer based feature fusion module, and use the simplest way to splice the features. It is used to verify the effectiveness of modal fusion.

The results of the ablation experiments are presented in Table IV.

Based on the experimental results, it's apparent that removing any module—be it the Modal Interaction Module, Local Similarity Module, or Feature Fusion Module results in a performance decline for the model. This underscores several significant findings:

Effective modal interaction facilitates the acquisition of more efficient features, thus enhancing overall model performance. This validates the efficacy of using word weight features to guide visual feature extraction.

The incorporation of improved Transformer encoding and decoding fusion helps in reducing information redundancy and noise interference, consequently leading to performance improvements.

Precise extraction of local similarity features plays a crucial role in mitigating graphical inconsistencies. Furthermore, it highlights the effectiveness of employing word weight features for fine-grained similarity computations within subgraphs.

TABLE IV. COMPARISON OF RESULTS OF ABLATION EXPERIMENTS

Datasets	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
WeiboA	CFIF-M	0.871	0.850	0.886	0.878	0.894	0.860	**0.887**
	CFIF-L	0.867	0.820	0.892	0.874	**0.920**	0.822	0.871
	CFIF-F	0.866	0.843	0.851	0.847	0.828	0.819	0.824
	CFIF	**0.882**	**0.883**	**0.901**	**0.881**	0.891	**0.873**	0.884
Twitter	CFIF-M	0.860	0.831	0.923	0.872	0.896	0.784	0.835
	CFIF-L	0.883	0.842	**0.962**	0.900	**0.952**	0.794	0.863
	CFIF-F	0.879	**0.881**	0.902	0.890	0.881	**0.862**	0.871
	CFIF	**0.890**	0.871	0.940	**0.901**	0.921	0.833	**0.872**

C. Parameter Analysis

In this study, we experiment and analyze two significant parameters with respect to two evaluation metrics: accuracy and F1 value. One parameter examines the impact of the number of heads on model performance within the modal fusion segment utilizing the multi-head attention mechanism. The other parameter investigates the effect of the number of output hidden layer neurons on model performance. The results of these analyses are presented in Fig. 7 and Fig. 8.

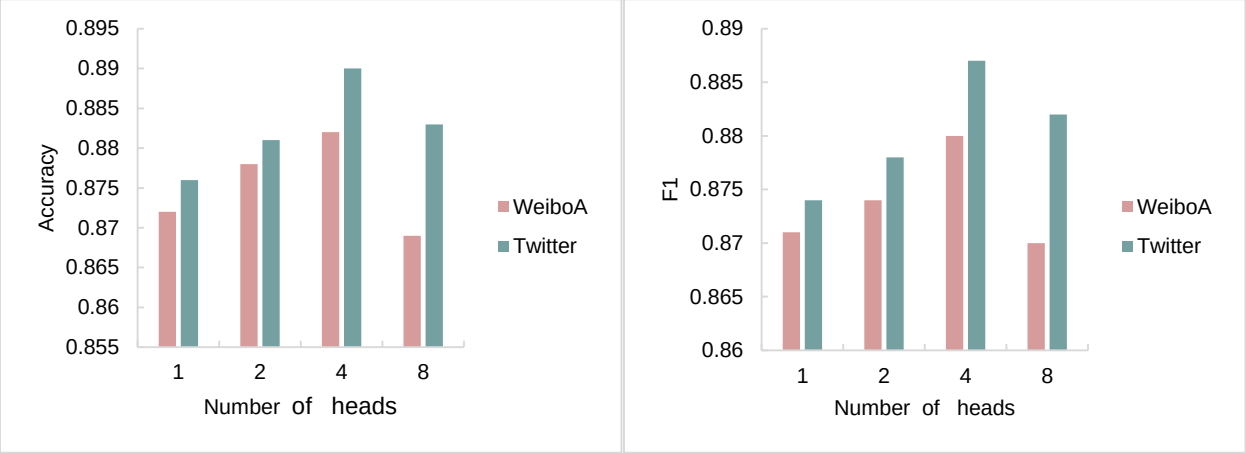


Fig. 7. The effect of different numbers of head on the results.

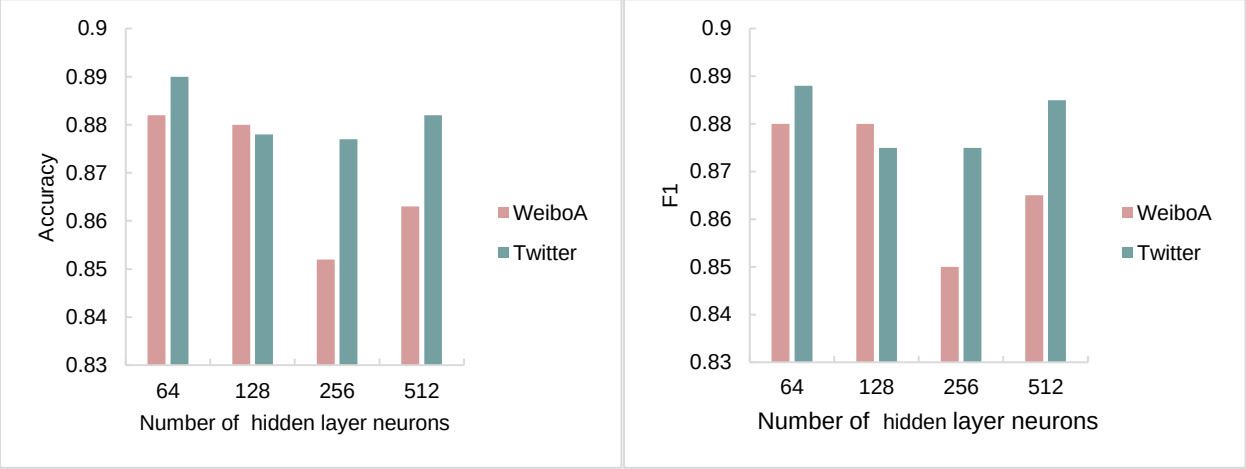


Fig. 8. The effect of different numbers of hidden layer neurons on the results.

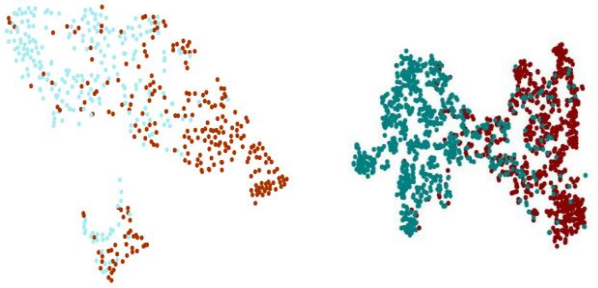
Based on the experimental findings, it's evident that an increased number of heads doesn't necessarily yield superior results. This phenomenon arises due to various factors such as computational resource constraints, overfitting, information redundancy, and challenges in hyperparameter selection. Firstly, augmenting the number of attention heads significantly escalates both the model's parameter count and computational complexity. This presents challenges in effective learning and optimization, particularly when resources are limited. Secondly, an excessive number of attention heads heightens the

risk of overfitting the training data, thereby reducing the model's ability to generalize to new data. Moreover, the information learned across attention heads may exhibit similarity or redundancy rather than complementarity, thereby diminishing the model's expressive power. Conversely, an increase in hidden layer neurons correlates with a decrease in accuracy due to neural networks with excessive hidden layer neurons being prone to overfitting. This abundance of nodes prolongs training time, hindering the achievement of desired outcomes. Hence, this study opts for optimal model performance, selecting four polyheads and 64 output hidden neurons as the preferred configuration.

D. Visualization Analysis

In order to further demonstrate the superiority of this paper's model, the dataset WeiboA is used as an example. The multimodal feature distribution of the test set is employed to visualize and analyze the classical fake news detection model EANN and this paper's model (CFIF). However, due to the high dimensionality of the multimodal features, it is challenging to intuitively understand the results. Therefore, the t-SNE algorithm is applied to map the multimodal feature dimensions to a two-dimensional space for visualization. The results are shown below in Fig. 9 multimodal feature distribution.

As can be seen from Fig. 9, for most of the data, both models are able to extract different features of real news and fake news, for the classification results, the more efficient the model, the closer the same class will be, and vice versa, the further away, while the EANN model is relatively loose regardless of the same class or different class spacing, which indicates that the uniqueness of the class features extracted by the model is relatively small, which is prone to lead to a lower performance of the model, and at the same time resulting in low generalization ability of the model. On the contrary, the model in this paper is able to make the news of the same category aggregated with small intervals on a great part of the data, while the different categories have large intervals at the same time, which reflects the importance of fine-grained modal interactions and good modal fusion.



(a) EANN feature distribution (b) CFIF feature distribution

Fig. 9. Multimodal feature distribution.

E. Fault Case Study

This subsection delves into the Chinese dataset WeiboA, with a focus on instances of model classification errors. Through an in-depth analysis of typical samples, we aim to discern the underlying reasons for these errors. In Fig. 10, false news is erroneously classified as true news. The textual content narrates activities related to visiting and traveling in Australia, accompanied by an image depicting similar activities. The convergence of textual and visual content makes it challenging to discern the veracity of the news solely based on internal cues. Consequently, our model identifies it as real news. In Fig. 11, real news is misclassified as false news. The textual content describes a father's affection for his son, whereas the accompanying image merely portrays an elderly father cooking. This discrepancy between the textual and visual elements results in a lack of coherence, leading our model to classify it as false news.

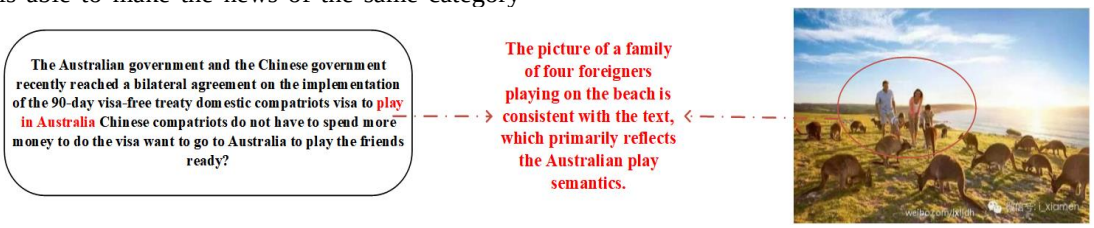


Fig. 10. Fake news judged to be true.

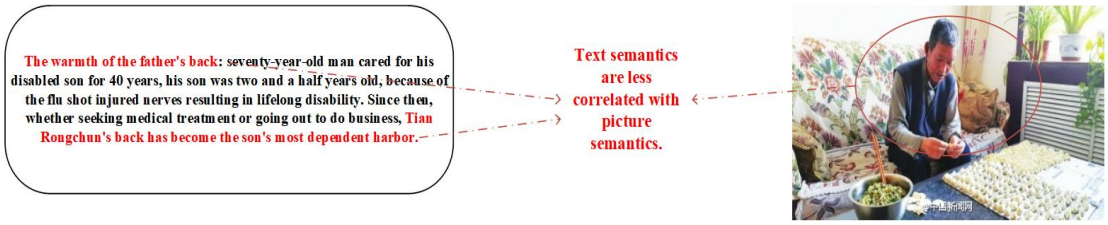


Fig. 11. True news judged to be fake.

VI. Conclusion

The proliferation of social media in recent years has facilitated easier access to information; however, it has also become a fertile ground for the propagation of false news. To enhance the efficacy of fake news detection, this paper proposes a cross-modal fine-grained interactive fusion model, primarily addressing the current issues of insufficient interaction between modalities and overly simplistic modality fusion in some detection models. This model achieves effective interaction by employing word weight features to guide the generation of visual features. Subsequently, it utilizes these features to compute fine-grained similarity across different regions of the

image, obtaining local similarity features. Finally, a fusion network is employed to integrate modal information, effectively mitigating the challenges associated with false news detection. Comparison experiments and ablation studies conducted on WeiboA and Twitter datasets demonstrate the efficacy of the proposed model compared to several benchmark models such as EANN.

Furthermore, the fine-grained fake news detection model proposed in this paper has some limitations, such as utilizing only part of the information within the dataset. Future studies can incorporate the remaining information as well as external data, such as the dissemination path of the news, user characteristics, and a priori characteristics combined with external information. Additionally, given the continuous evolution of technology and the sophistication of tampering methods, a significant proportion of fake news content falls into a gray area, blending elements of truth and falsehood. Therefore, future efforts should avoid oversimplifying the fake news detection task as a binary classification problem and instead develop it into a multi-classification challenge.

ACKNOWLEDGMENT

Henan Provincial Science and Technology Plan Project (No: 212102210417).

REFERENCES

[1] X. Tang , C. Huang & X. Wu. “China New Media Development Report No. 11,”Social Science Literature Press, Beijing,2020.

[2] P. Meel & D. K. Vishwakarma, “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,” Expert Systems with Applications, vol. 153, p. 112986, Sep . 2020.

[3] X. Zhou , R. Zafarani, K. Shu & H. Liu, “Fake news: Fundamental theories, detection strategies and challenges,” InProceedings of the twelfth ACM international conference on web search , pp. 836-837, 2019.

[4] B. Horne & A. Sibel, “This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news,” In Proceedings of the international AAAI conference on web and social media, pp. 759-766, 2017.

[5] V. Pérez-Rosas, B. Kleinberg & A. Lefevre, “Automatic detection of fake news,”Proceedings of the 27th International Conference on Computational Linguistics, New Mexico, USA, pp. 1348-2021, 2018.

[6] C. Castillo, M. MendozaE & B. Poblete, “Information credibility on twitter,”Proceedings of the 20th International Conference on World Wide Web, pp. 675-684,2011.

[7] Z. W. Jin, J. Cao. B. Wang, R. Wang & Y. D. Zhang, "Rumor detection on social media with multimodal feature fusion," Journal of Nanjing University of Information Science and Technology, vol.9, no.6, pp. 583-592, 2017.

[8] C. Y. Chen & J. Sui, “An integrated multimodal rumor detection method based on DeepFM and convolutional neural network,” Computer Science, pp. 101-107, 2020.

[9] Y. Q. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection" Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, London, UK, pp. 849-857, 2018.

[10] J. Ma, W. Gao, & P. Mitra, “Detecting rumors from microblogs with recurrent neural networks,”Proceedings of the Twenty- Fifth International Joint Conference on Artificial Intelligence, pp. 3818-3824, 2016.

[11] S. Volkova, K. Shaffer & J. Y. Jang, “Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter,”Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, pp.647-653, 2017.

[12] S. Chawda, A. Patil & A. Singh, “A Novel Approach for Clickbait Detection,”Proceedings of 2019 3rd International Conference on Trends in Electronics and Informatics, India, pp.1318-1321, 2019.

[13] H. He & R.Wang. “A fake news content detection model based on feature aggregation,” Computer Science, pp. 1 -7, 2020.

[14] P. Qi, J. Cao & T. Yang, “Exploiting multi-domain visual information for fake news detection,” IEEE International Conference on Data Mining（ICDM）, pp. 518 -527, 2019.

[15] J. Xue, Y. Wang & S. Xu, “Mvfnn:multi- vision fusion neural network for fake news picture detection,” International Conference on Computer Animation and Social Agents, Cham,Springer,pp. 112 -119, 2020.

[16] P. Zhou, X. Han, V. Morariu & L. S. Davis,“Learning rich features for image manipulation detection,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1053-1061, 2018.

[17] X. Y. Zhou, J. D. Wu & R. Zafarani, “SAFE: Similarity-Aware Multimodal Fake News Detection,” In Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, Singapore, pp.354–367, 2020.

[18] C. Song, N. Ning & Y. Zhang, “A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural net-works,” Information Processing & Management, p. 102437, 2021.

[19] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, 2019.

[20] T. He, W. Huang, Y. Qiao & J.Yao, “Text-attentional convolutional neural network for scene text detection,” IEEE transactions on image processing,pp. 2529-2541, 2016.

[21] Y. Liu, H. Liu & L. P. Wong, “A hybrid neural network RBERT-C based on pre-trained RoBERTa and CNN for user intent classification,” Neural Computing for Advanced Applications: First International Conference, NCAA 2020, Shenzhen, China,pp. 306-319, 2020.

[22] T. Mikolov, K Chen, G Corrado & J Dean, “Efficient estimation of word representations in vector space,” 2013.

[23] T. A. Jibril & M. H. Abdullah. “Relevance of Emoticons in ComputerMediated Communication Contexts: An Overview,” Asian Social Ence, pp. 201-207, 2013.

[24] J. Yoon & E. Chung, “Image Use in Social Network Communication: A Case Study of Tweets on the Boston Marathon Bombing,” Information Research, pp. 106-116, 2016.

[25] K. He, X. Zhang & S. Ren, “Deep Residual Learning for Image Recognition,” Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016.

[26] Z. Jin, J. Cao, H. Guo, Y. Zhang & J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," Proceedings of the 25th ACM international conference on Multimedia, New York, USA, pp. 795-816, 2017.

[27] C. Boididou, et al, "Verifying multimedia use at mediaeval 2015," MediaEval, vol. 3, no. 3, pp. 7, 2015.

[28] D. Khattar, J. S. Goud, M. Gupta & V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," The world wide web conference, San Francisco, CA, USA, pp. 2915-2921, 2019.

[29] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi & L. Wei, “Detecting fake news by exploring the consistency of multimodal data,” Information Processing & Management, no. 5, p.102610, 2021.

[30] B. Singh & D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," Neural Computing and Applications, vol. 34, no. 24, pp. 21503-21517, 2022.

[31] P. Wei, F. Wu, Y. Sun, H. Zhou & X. Y. Jing, "Modality and event adversarial networks for multi-modal fake news detection," IEEE Signal Processing Letters, vol. 29, pp.1382-1386, 2022.