Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053937

ARTICLE

Fake News Detection Based on Cross-Modal Message Aggregation and Gated Fusion Network

Fangfang Shan[1,2,*], Mengyao Liu[1,2], Menghan Zhang[1,2] and Zhenyu Wang[1,2]

[1]College of Computer, Zhongyuan University of Technology, Zhengzhou, 450007, China

[2]Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou, 450001, China

*Corresponding Author: Fangfang Shan. Email: 6129@zut.edu.cn

Received: 13 May 2024 Accepted: 12 June 2024 Published: 18 July 2024

ABSTRACT

Social media has become increasingly significant in modern society, but it has also turned into a breeding ground for the propagation of misleading information, potentially causing a detrimental impact on public opinion and daily life. Compared to pure text content, multmodal content significantly increases the visibility and share ability of posts. This has made the search for efficient modality representations and cross-modal information interaction methods a key focus in the field of multimodal fake news detection. To effectively address the critical challenge of accurately detecting fake news on social media, this paper proposes a fake news detection model based on cross-modal message aggregation and a gated fusion network (MAGF). MAGF first uses BERT to extract cumulative textual feature representations and word-level features, applies Faster Region-based Convolutional Neural Network (Faster R-CNN) to obtain image objects, and leverages ResNet-50 and Visual Geometry Group-19 (VGG-19) to obtain image region features and global features. The image region features and word-level text features are then projected into a low-dimensional space to calculate a text-image affinity matrix for cross-modal message aggregation. The gated fusion network combines text and image region features to obtain adaptively aggregated features. The interaction matrix is derived through an attention mechanism and further integrated with global image features using a co-attention mechanism to produce multimodal representations. Finally, these fused features are fed into a classifier for news categorization. Experiments were conducted on two public datasets, Twitter and Weibo. Results show that the proposed model achieves accuracy rates of 91.8% and 88.7% on the two datasets, respectively, significantly outperforming traditional unimodal and existing multimodal models.

KEYWORDS

Fake news detection; cross-modal message aggregation; gate fusion network; co-attention mechanism; multi-modal representation

1 Introduction

The proliferation of fake news has become a grave concern in light of the Internet and social media's rapid advancement. The number of internet users continues to grow, and the number of online news consumers is also on the rise. However, because people often share information without thoroughly checking its authenticity, and because social media, unlike traditional media, allows for

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

faster and broader dissemination, the resulting negative impacts can be significant, with even greater harm to society. This problem became particularly acute during the COVID-19 pandemic when the internet was flooded with fabricated fake news, such as claims that tea, vinegar, and saltwater could effectively treat COVID-19 [1]. The spread of such misinformation poses a tremendous risk to public health and safety. Thus, effectively detecting and stopping the fake news spreading on social media has become an urgent social issue. Addressing this problem is crucial for purifying the online environment, maintaining social stability, and ensuring national information security.

Early methods for fake news detection primarily relied on manual fact-checking and expert judgment, which were time-consuming, labor-intensive, and inefficient. Traditional machine learning approaches employed algorithms such as Naive Bayes [2], decision trees [3], or support vector machines [4] to classify extracted text features. However, feature selection directly influenced the performance of these classifiers, and they struggled to accommodate the diversity of news content.

Social media platforms like Weibo have rapidly evolved from hosting purely text-based content to featuring multimedia forms, including text and images. While initial endeavors to detect fake news primarily concentrated on analyzing textual content, cross-modal content analysis can provide supplementary benefits and aid in fake news detection. With the development of deep learning, researchers have begun using deep neural networks to learn higher-level feature representations from news, such as employing pre-trained convolutional neural networks (CNN) and variational auto-encoders (VAE) [5] to extract multimodal features. These features are then combined with different multimodal feature fusion methods to construct end-to-end multimodal detection frameworks. Despite the effectiveness of these existing fake news detection approaches, they still have some limitations: First, traditional machine learning and deep learning models may lose some critical information during the feature extraction process, potentially limiting the model's performance and detection accuracy. Second, existing multimodal feature fusion methods might lead to the accumulation of irrelevant information across different modalities, generating excessive redundant data can diminish the efficiency of the model.

Current research on multimodal fake news detection, both domestically and internationally, primarily focuses on simple fusion analysis of news text and images without fully exploring the similarity and degree of association between different modalities. The interaction between text and image information is often shallow, failing to take full advantage of multimodal semantic features. For example, in Fig. 1, the news text describes a death story, while the accompanying image shows a smiling person, indicating a high level of cross-modal ambiguity between the text and image. For this situation, cross-modal information association and the aggregation and transmission of local messages can capture the gaps between cross-modal information, helping to improve the accuracy of news classification. By effectively integrating cross-modal data, not only can the fine-grained associations between text and images be revealed, but the complementary information from both modalities can also be utilized to enhance the ability to detect fake news.

In order to address the limitations and issues mentioned above, this study introduces a novel approach for detecting fake news using a multimodal framework that incorporates cross-modal message aggregation and gated fusion. The approach consists of five modules: (1) a multimodal feature extractor; (2) cross-modal adaptive message passing based on association matrices; (3) a gated fusion network; (4) multimodal feature fusion; and (5) a fake news detector. Most previous methods only focus on global or local features of text and images. However, we comprehensively consider both local and global features during feature extraction. When people read news with images and text, they usually alternate their attention between the sentences in the news text and the related image

regions, considering the interaction between these two modalities. Inspired by this, we introduced a Cross-Modal Message Aggregation module into our model. This module focuses on salient regions in images and keywords in news text, exploring the subtle associations between text and images, and extracting fine-grained clues. This enables effective integration of information interaction between text and images, thereby enhancing the comprehensive judgment of news authenticity. Cross-modal data leverage the complementary dependencies between modalities to select information that is beneficial for identifying fake news. Thus, our approach takes into account detailed cross-modal interactions and introduces a gated fusion network. This network uses an adaptive gating mechanism to appropriately handle task-irrelevant information. In fake news detection, finding and integrating fine-grained cross-modal clues is crucial, and task-irrelevant information should be suppressed during information transfer. Hence, we employ a soft gated fusion mechanism to adaptively control the degree to which messages from one modality are fused with the original features of the other modality.

Fake news: "An employee of the Jefferson County morgue died this morning after being accidentally cremated by one of his coworkers."

Figure 1: Example of fake news

The primary contributions of this paper are summarized as follows:

(1) This paper considers the local features of both text and images, using cross-modal message aggregation to account for bidirectional information passing between word domains. This enables the model to investigate detailed interactions between different modalities, thereby effectively identifying fake news.
(2) For text, the model employs a gated fusion network to optimize feature fusion. Through a soft gating mechanism, it can adaptively control the intensity of information fusion, facilitating more profound and exhaustive interactions between news images and text. This adaptive gating helps manage the effects of negation pairs and irrelevant background information, as a result, this enhances the precision and effectiveness of fake news detection.
(3) The performance of the proposed model is assessed using two extensive real-world datasets. Results indicate that compared to traditional detection models, the proposed model demonstrates superior performance across various evaluation metrics, leading to a substantial enhancement in the accuracy and overall performance of fake news detection.

2 Related Work

Fake news can be defined as deliberately written information that can be proven to be false [6]. Given the significant harm that fake news can cause, fake news detection has been a widely discussed research area. Fake news detection methods currently in use can be classified into two main types: unimodal and multimodal approaches.

2.1 Unimodal-Based Fake News Detection

The efficacy of fake news detection largely depends on the analysis of text content. As one of the primary mediums for spreading fake news, textual features such as vocabulary, grammar, and semantics [7–9] can offer import clues for discerning the credibility of news articles. Consequently, many early researchers focused on developing and applying techniques like machine learning and deep learning to extract text features. Statistical-based methods [10,11] rely on statistical information within the text, such as term frequency, inverse document frequency, and term frequency-inverse document frequency, to extract significant text features for classification and analysis. These statistical features can reveal patterns, trends, or key information within the text, aiding in distinguishing fake news. For example, Liu proposed extracting text feature vectors through the term frequency-inverse document frequency (TF-IDF) algorithm and using a support vector machine (SVM) to determine the authenticity of news stories [12].

In recent years, researchers have begun using pre-trained language models to enhance the performance of fake news detection, which can enable better handling of various natural language tasks. These methods have significantly impacted fake news detection. For example, the Bidirectional Encoder Representations from Transformers (BERT) [13] model has been applied to fake news detection, yielding better performance compared to traditional machine learning models. Neural networks have also been used in fake news detection tasks. Chen et al. [14] proposed a deep attention model based on Recurrent Neural Networks (RNN) to identify fake news. This model deeply studies the repeated use of soft attention to simultaneously focus on distinct features with specific foci and produce hidden representations to capture contextual variations in news content over time. Ma et al. [15] employed Generative Adversarial Networks (GAN) to introduce stronger text representation learning.

With the rise of multimodal news content, researchers are increasingly focusing on leveraging visual features to detect fake news. Mahmood et al. [16] combined Stationary Wavelet Transform with Discrete Cosine Transform (DCT) to detect and locate copy-move operations in images. Qi et al. [17] proposed the MVNN (Multi-domain Visual Neural Network) framework for detecting fake news images. This framework extracts physical-level features from images and processes and analyzes images spatially to extract their pixel-domain features. The features are then dynamically fused through a fusion sub-network, resulting in a model that outperforms traditional models. Although visual features can provide rich visual information, the current comprehension of the significance of visual features in fake news detection still has certain limitations [18].

2.2 Multimodal-BasedFake News Detection

In recent years, with the increasing diversification of fake news articles, traditional unimodal feature approaches have become inadequate for fully leveraging multimodal feature information to distinguish between real and fake news. To better detect the authenticity of news, there is a need to deeply integrate text and image features to enhance fake news detection performance. A common approach is to use different feature extractors to process text and images, then concatenate the extracted features from various modalities, and classify them using different classifiers. Jin et al. [19] proposed a modal that integrates multimodal features using an attention mechanism. Wang et al. [20] introduced the Event Adversarial Neural Network (EANN) for effectively detecting fake news and accurately discriminating events. The framework includes an event discriminator to eliminate event-specific features. Singhal et al. [21] proposed a framework called SpotFake. The model captures textual features from news articles using the BERT, while visual features are extracted from VGG-19 model on the ImageNet dataset, improving the performance of fake news detection. Qian et al. [22] proposed a

Hierarchical Multimodal Context Attention Network (HMCAN), which inputs multimodal features into a multimodal context attention network for joint modeling to integrate inter- and intra-modal relationships. Wu et al. [23] proposed a Multimodal Co-Attention Network (MCAN), where CNN and VGG-19 are used to capture visual representations at the physical and semantic levels, while BERT is used to capture text representations. Multiple stacked co-attention blocks are used to fuse these modalities, allowing MCAN to capture the dependencies between modalities. Shan et al. [24] proposed employing similarity reasoning and adversarial networks, using similarity learning and reasoning to create similarity representations between textual and visual features, and using adversarial networks to explore the relationship between fake news and events.

Multimodal fake news detection seeks to improve detection performance through deep integration of text and image features. However, challenges persist, such as insufficient correlation between data features from different modalities and inadequate modality interaction. To tackle these challenges, we propose a method for fake news detection based on cross-modal message aggregation and gated fusion.

3 Method

3.1 Modal Overview

This paper proposes MAGF, a method that addresses multimodal fake news detection through cross-modal message aggregation and gated fusion. As shown in Fig. 2, MAGF consists of the following components: 1) Multimodal Feature Extractor, which extracts text features using BERT, employs a pre-trained Faster R-CNN [25] to obtain image region features, and uses VGG-19 to obtain image features from the entire image; 2) Cross-Modal Message Aggregation, which calculates a text-image association matrix using text features and image region features, subsequently performs message passing to acquire aggregated features for each modality; 3) Gated Fusion Network, which computes gating values for each region feature based on the image region features and the aggregated text features from cross-modal message aggregation, allowing control over how much information should be transmitted during cross-modal fusion. This same gating process is applied to the original text features; 4) Multimodal Feature Fusion, which utilizes a simple attention mechanism to integrate the gated features into an interaction matrix and then combines it with global image features to acquire the ultimate feature representation; 5) Fake News Detector, which uses the multimodal representation to make the final prediction for the news article.

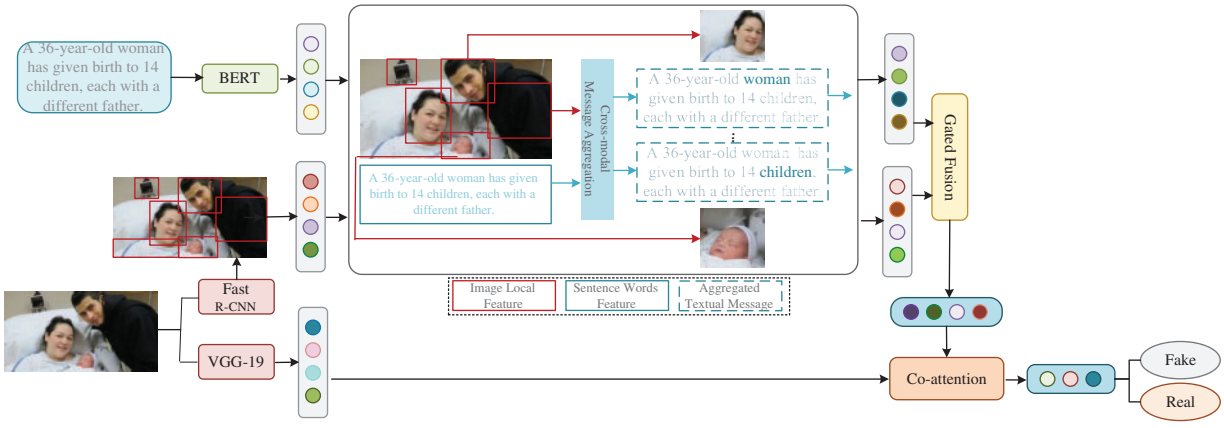


Figure 2: Framework of MAGF

3.2 Multimodal Feature Extractor

3.2.1 Textual Feature Extractor

This paper uses the BERT model as the text feature extractor, as BERT is designed to better understand the semantic relationships between words and the context. Given a text sentence C from the news article, a simple processing step creates an ordered list of words, and then context information is extracted from the input text to obtain both global and word-level features. Specifically, we utilizes the [CLS] token and token-level output embeddings. In this process, the [CLS] token represents the semantic information of the entire sentence, serving as the "global representation" of the sentence, while the token-level output embeddings represent the specific semantic information of each token within its context. By combining these two types of representations, it is possible to acquire a cumulative feature representation for the entire text, as well as feature representations at the word level. The text feature T is represented as shown in Eq. (1).

$$T = \{t_0, t_1, t_2, \ldots t_r\}, t_i = BERT\,(C, w_i), \forall w_i \in \{w_1, w_2, \ldots, w_r\} \tag{1}$$

where t_i is a 256-dimensional feature vector, and t_0 is the sentence-level representation.

3.2.2 Visual Feature Extractor

For news images, it is necessary to extract both image region features and global features. Image region features capture information from each local region of the image, providing finer-grained visual details, including local objects, textures, colors, and more. The combination of the two allows for a more effective use of semantic information and fine-grained features from the image.

(1) Image region features

Faster R-CNN is used to extract candidate regions from an image, potentially containing objects (known as bounding boxes). It employs a sliding window mechanism on the feature map to extract fixed-size anchor boxes and predicts two values for each anchor: adjustments to the bounding box coordinates and the confidence score for the target object. These predictions are then used to filter and rank the candidate bounding boxes, selecting those with high confidence scores as the final candidate regions. By using these bounding boxes to crop out local areas that contain objects, the regions of interest can be isolated from the entire image. This helps improve the efficiency and accuracy of subsequent feature extraction.Specifically, we use Faster R-CNN model to obtain the bounding boxes $G = \{gg_1, gg_2, \ldots, gg_n\}$ from a news image, where gg_1 represents the i-th object detected by the model. The bounding boxes are then used to crop the image to obtain region-specific objects, which are fed into the pre-trained ResNet-50 [26] model to obtain image region features I, it can be represented as shown in Eq. (2).

$$I = \{v_1, v_2, \ldots v_n\}, v_i = ResNet\,(gg_i), \forall gg_i \in \{gg_1, gg_2, \ldots gg_n\} \tag{2}$$

(2) Image global features

VGG-19 is utilized to extract the global features, as its deep architecture allows the network to gradually extract hierarchical features from the image. Furthermore, VGG-19 has achieved good performance in the ImageNet image classification competition, demonstrating its wide applicability in various domains. Therefore, using VGG-19 to extract global features can better understand the entire content of the image. The obtained image global features are represented by Eq. (3).

$$V = VGG - 19\,(img) \tag{3}$$

3.3 Cross-Modal Message Aggreation

Many existing methods often fail to avoid generating a large amount of redundant data when processing multimodal information, while easily overlooking the critical information that needs to be transmitted between different modalities. This insufficiency in information processing may result in a decrease in the modal's accuracy of detection. Therefore, in the model designed in this paper, a cross-modal message aggregation module is specifically introduced. This module aims to extract key information from each modality and effectively aggregate information between regions and words, thus gaining a deeper understanding of the intrinsic relationship between text and images. At the same time, it reinforces crucial information exchange across diverse modalities, thereby enhancing the model's efficacy.

(1) Region-Word Affinity Matrix

To aggregate messages between text and vision, the first step is to calculate the region-word affinity matrix, which provides a tool for the model to quantify the correlation between different modalities. First, the extracted word-level text features T and the regional features of the image I are mapped into a unified low-dimensional feature space. In this space, the similarity between each regional feature vector and each word feature vector is measured, and then the region-word affinity matrix is constructed. The rows of the matrix correspond to different regions in the image, while the columns correspond to individual words in the text. The matrix can be represented by Eq. (4).

$$\mathrm{S} = (W_v I)^{\mathrm{T}} (W_t T) \tag{4}$$

where W_v and W_t are projection matrices, $S \in R^{R\times \mathrm{N}}$ is the region-word affinity matrix, where S_{ij} represents the affinity between the i-th region and the j-th word. This transformation effectively quantifies the degree of association between image regions and textual words.

(2) Attention Guidance

Our goal is to effectively aggregate and propagate key information between different information regions and textual words. Therefore, we utilize a cross-modal attention mechanism that leverages information from two modalities (such as text and images) as cues to guide the model to focus more precisely on critical information within the same modality. Specifically, at the word level, we utilize the features of image regions as guidance to assign attention weights to each word in the text. Through this approach, the model can filter out word information that is closely related to the image content, achieving effective alignment between text and images at the word level. Simultaneously, at the regional level, we also utilize the features of textual words as cues to assign attention weights to various regions in the image. This helps the model capture image information that matches the textual content, achieving precise correspondence between images and texts at the regional level.

The attention for each word on different regions is obtained by normalizing the affinity matrix across image regions, resulting in a word-specific region attention matrix, which can be represented by Eq. (5).

$$S_I = soft\max\left(\frac{S^{\mathrm{T}}}{\sqrt{d_h}}\right) \tag{5}$$

where the i-th row represents the degree of attention on all regions regarding the i-th word. Then, based on the word-specific regional attention matrix S_I, the regional features R_I for each word can be obtained, which can be represented by Eq. (6).

$$R_I = S_I I^{\mathrm{T}} \tag{6}$$

The i-th row of $R_I \in R^{N\times d}$ represents the image region that accompany the i-th word.

Correspondingly, by normalizing the affinity matrix S at the word level, the attention weights of each word relative to each image region can be precisely calculated, resulting in a word attention matrix S_T. Then the text local features are aggregated to get the text features T_t that each region focuses on. This can be represented by Eq. (7).

$$S_T = soft\max\left(\frac{S}{\sqrt{d_h}}\right), T_t = S_t T^{\mathrm{T}} \tag{7}$$

In summary, R_I and T_t refer to the aggregated messages that are intended to be exchanged between visual local features and textual features. This approach allows for a deeper understanding of the correlation between text and images and better realization of cross-modal interactions.

3.4 Gated Fusion Network

While the aforementioned modules integrate the most significant cross-modal messages from word or region features, thereby facilitating the exchange of information between the text and visual modalities, the final step in fake news detection requires multi-modal fusion of these features. To achieve this, we need an efficient fusion process that strengthens the fusion of matching text and visual features while suppressing the fusion of non-matching pairs, ensuring accurate information transfer and in-depth interaction. For this purpose, a cross-modal gated fusion network was designed, as depicted in Fig. 3. This mechanism introduces a gating unit to control the degree of fusion between text and image features, enabling precise matching of feature fusion while avoiding interference from non-matching information.

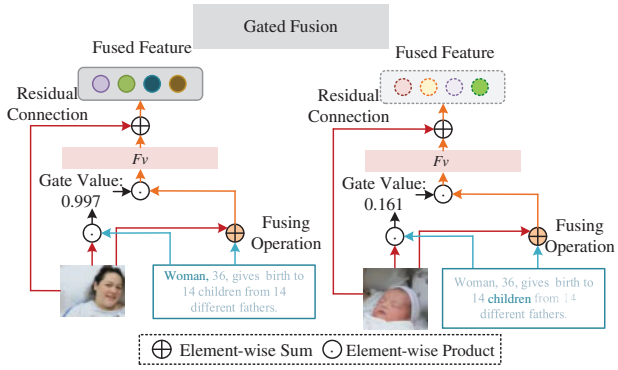


Figure 3: Gated fusion network

First, calculate the gating values using the messages transmitted from visual region features and text features separately. Then, fuse the original region features I with the aggregated message T_t using gating units. To filter out non-matching information during fusion, adaptive control of the information fusion is applied at the region-word level. Specifically, the gating value (i.e., the matching degree) corresponding to the i-th region feature $\mathrm{v_i}$ and the corresponding sentence feature $m_i{}^t(\in T_t)$ that the i-th region is focusing on can be represented by Eq. (8).

$$x_i = \sigma\left(v_i \odot m_i{}^t\right), i \in \{1,\ldots,n\}\, m_i{}^t\ (\in T_t) \tag{8}$$

where $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function, and x_i represents the gate that fuses v_i and $m_i{}^t$. With this gating function, If a region demonstrates a good match with a

sentence, it will be assigned a high gating value, thereby promoting the fusion operation. Conversely, if a region does not exhibit a good match or alignment with a sentence, it will be assigned a low gating value, leading to the suppression of the fusion operation. All the region-level gating values are then represented as $X_v = [x_1, ..., x_n]$, which will be used to control cross-modal fusion. To prevent the loss of important information from the original samples, we integrate the information processed through gating units with the original features, which can be represented by Eq. (9).

$$R_I' = F_v(X_v \odot (I \oplus T_t)) + I \tag{9}$$

where F_v is a transformation operation adjustable through learning. $\odot$ represents hadamard product, $\oplus$ denotes fusion (component-wise addition), while $R_I^{'}$ represents the fused region features. Regions that match well with sentences are assigned higher gating values to encourage deeper fusion, while for negative pairs with lower gating values, the gating mechanism suppresses the fusion, thereby encouraging $R_I^{'}$ to retain its original features, V. Similarly, the original text features, T, and the aggregated message, R_I, transmitted from region features to text features are fused to obtain the fused text features, T_t'. This is computed according to Eqs. (10)–(12), where $m_i^v \in R_I$.

$$z_i = \sigma\left((m_i^v)^{\mathrm{T}} \odot t_i\right), i \in \{1, \ldots, r\} \tag{10}$$

$$Z_t = [z_i, \ldots, z_r] \tag{11}$$

$$T_t' = F_t\left(Z_t \odot \left(T \oplus R_I^{\mathrm{T}}\right)\right) + T \tag{12}$$

3.5 Multimodal Feature Fusion

Based on the gated fusion network, after obtaining $R_I^{'}$ and $T_t^{'}$, we integrate them using a simple attention mechanism. Specifically, through linear projection and softmax normalization, we compute the attention weight matrix and then aggregate the given $R_I^{'}$ and $T_t^{'}$ with the attention weights to get R_I^* and T_t^*, as represented by equations Eqs. (13), (14).

$$\alpha = soft\max\left(\frac{W_I R_I'}{\sqrt{d}}\right)^{\mathrm{T}}, R_I^* = R_I\alpha \tag{13}$$

$$\beta = soft\max\left(\frac{W_T T t'}{\sqrt{d}}\right), T_t^* = T_t\beta \tag{14}$$

Here, W_I and W_T are the parameters for linear projections, while α and β are the attention weights for the fusion of n regions and r words, respectively. The above formulas combine image region features with text features. The image region features can capture local details and key information within images. These region features help the model understand the relationship between specific parts of the image and the text content. However, solely using image region features for news detection might lead to partial interpretations. Therefore, it's also necessary to incorporate image global features, which provide an overview of the entire image,offering a higher degree of generalization.We first construct the interaction matrix F by computing the outer product between the aggregated R_I^* and T_t^*, represented by Eq. (15).

$$F = R_I^* \otimes T_t^* \tag{15}$$

The model uses a co-attention block [27] to fuse the global image features with R_F. Its structure is shown in Fig. 4. In the co-attention block, one modality's features are used as queries, while the other modality's features serve as keys and values. This setup can enhance the expressiveness of each

modality's features, allowing for information exchange and interaction between different modalities, thereby improving the overall quality and richness of feature representation. The final multimodal feature representation, R_H, is obtained by Eqs. (16)–(20).

$$Att\left(Q,K,V\right)=hW^{e},h=h_1\oplus h_2\oplus\ldots\oplus h_m \tag{16}$$

$$h_i=A\left(Q_i,K_i,V_i\right)=soft\max\left(\frac{Q_iK_i^{\mathrm{T}}}{\sqrt{d_h}}\right)V_i$$

$$LRF\left(u\right)=\max\left(0,uW_1\right)W_2 \tag{17}$$

$$R_{H_{F\leftarrow V}}=F+Att\left(F,V,V\right),R_{H'_{F\leftarrow V}}=R_{H_{F\leftarrow V}}+LRF\left(R_{H_{F\leftarrow V}}\right) \tag{18}$$

$$R_{H_{V\leftarrow F}}=V+Att\left(V,F,F\right),R_{H'_{V\leftarrow F}}=R_{H_{V\leftarrow F}}+LRF\left(R_{H_{V\leftarrow F}}\right) \tag{19}$$

$$R_H=\left(R_{H'_{F\leftarrow V}}\oplus R_{H'_{V\leftarrow F}}\right)W_H \tag{20}$$

In Eq. (18), $R_{H'_{F\leftarrow V}}$ is the attention-pooled feature based on the interaction matrix with the global image features. While $R_{H'_{V\leftarrow F}}$ is the attention-pooled feature based on the global image features from the interaction matrix. W_H represents the learnable parameters.

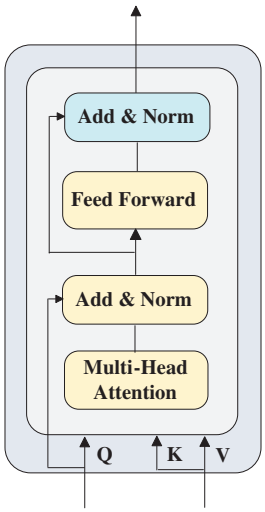


Figure 4: The structure of co-attention block

3.6 Fake News Detection

For this module, the obtained feature representation R_H needs to be input to the fully connected layer, and the output of the fully connected layer is passed through softmax to generate the distribution of classification labels The value of each dimension represents the probability of the input sample belonging to the corresponding class. Therefore, by comparing these probability values, one can determine the most likely class that the input news belongs to, i.e., real news or fake news. The fake news detector can be represented by Eq. (21).

$$\hat{y}=soft\max\left(\max\left(0,R_HW_{R_H}\right)W_s\right) \tag{21}$$

where $\hat{y}$ represents the probability distribution of the current news, R_H is the fused multimodal feature, The model's loss function utilizes minimizing the cross-entropy loss function, represented by Eq. (22).

$$L(\Theta) = -y \log\left(\hat{y}\right) - (1-y) \log\left(1-\hat{y}\right) \tag{22}$$

The true label of the news is represented by y, where a value of 1 indicates fake news and a value of 0 indicates real news, the Θ encompasses all the trainable parameters within the model.

4 Experiment

4.1 Dataset

Experiments were conducted on the Twitter and Weibo datasets. The Twitter dataset was released by MediaEval for multimedia usage tasks [28], while the Weibo dataset was collected by Jin et al. [19].

Twitter Dataset: The dataset is partitioned into two segments: a development set and a test set. The development set is utilized for training purposes, while the test set is used for evaluation during the experiments. The model in this article detects fake news by combining textual and image information. Therefore, we removed tweets without any text or images during data preprocessing.

Weibo Dataset: The genuine news in this dataset was verified by the authoritative Chinese news agency Xinhua, while the fake news was verified by Weibo's official rumor-debunking system. This means that the dataset contains not only real news but also verified fake news. Therefore, this dataset provides reliable and effective data support for fake news detection. Similarly, we removed news items with videos and those without text or images during data preprocessing to ensure the quality of the dataset.

The statistical data of the two datasets are shown in Table 1.

Table 1: Statistics of two datasets

Method	Twitter	Weibo
#Of fake news	7898	4211
#Of real news	6026	3639
#Of images	512	7850

4.2 Expreimental Setup

On the Weibo dataset, we set the maximum length of text to 160. For the Twitter dataset, we used the pre-trained "bert-base-uncased" model from HuggingFace 5, while we used the pre-trained "bert-base-chinese" model on Weibo. During the training phase, the parameters of the pre-trained VGG-19 and BERT networks are frozen to avoid overfitting. The text feature dimension obtained from the BERT model is 768. In the visual encoder, the input image size is 224 $\times$ 224, For computational convenience, an additional fully connected layer is added in the proposed model to ensure that the dimensions of image and text features match in the multimodal feature fusion module. Specifically, the dimension of the common embedding space where the final text and image features are projected is set to 256. During the feature fusion process, the number of attention heads in the multi-head self-attention mechanism is set to 4, which helps reduce computational costs and training time. Our proposed model adopts an early stopping strategy during training. For the Twitter and Weibo datasets,

we used Adam [29] and AdaBelief [30] optimizers respectively to optimize the model parameters. The other parameter configurations set in the model can be found in Table 2.

Table 2: The other parameter configurations of the model

Parameters	Value
Dropout	0.5
Batch size	32
Learning rate	0.001
Epoch	50
Early stopping patience	10

The article leverages multiple pre-trained submodels, such as BERT and VGG-19, to address the computational complexity and resource requirements. They demonstrate robust performance across various tasks and possess rich feature extraction capabilities. Training a full-depth model like VGG-19 from scratch demands significant GPU (graphics processing unit) resources and ample memory. Utilizing pre-trained models mitigates the need for training from scratch, thereby significantly reducing the demands on computational resources and time.

In the process of multimodal feature fusion, the model employs a co-attention mechanism. Compared to fully connected networks or large convolutional neural networks, attention mechanisms reduce the number of features to be processed by attending to relevant information, thereby reducing model parameters. This not only lowers memory requirements but also decreases computational complexity. The experimental setup for this study is outlined in Table 3.

Table 3: Experiment environment configuration

Experiment environment	Configuration
CPU	Intel(R) Core(TM) i9-12900K
GPU	NVIDIA RTX A5000
RAM (random access memory)	64
Programming language	Python 3.8
Pytorch version	1.8.2
CUDA toolkit	11.6

4.3 Evaluation Metrics

This paper utilizes F1 score, accuracy, precision and recall as evaluation metrics to quantify the performance of fake news detection. The following provides explanations for these metrics:

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN} \tag{23}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

$$F_1 = \frac{2 * \Pr ecision * Recall}{\Pr ecision + Recall} \tag{26}$$

where True positive (TP): Fake news forecasted as fake; True negative (TN): Real news forecasted as real; False positive (FP): Real news forecasted as fake; False negative (FN): Fake news forecasted as real.

4.4 Baseline

4.4.1 Unimodal Models

In this model, either textual or visual information is used alone to detect fake news. Therefore, this paper proposes the following two baselines:

• **Textual.** This model only uses the textual information in the news for classification. It first obtains textual features through the BERT model, and then uses another fully connected layer with a softmax function to detect fake news.

• **Visual.** This model utilizes only the visual information in the news to classify the posts. It first extracts image features using a pre-trained VGG-19 model and a fully connected layer, and then inputs them into another fully connected layer for fake news detection.

4.4.2 Multimodal Models

• **EANN (Event adversarial neural network) [20].** This model introduces an adversarial neural network for the auxiliary task of removing event-specific features. In the original datasets, the event labels are not included. To ensure a fair comparison, we remove the event labels from the datasets.

• **MVAE (Multimodal variational autoencoder) [5].** This model combines a bimodal VAE with a binary classifier to detect fake news by learning shared representations of textual and visual information.

• **MPFN (Multimodal progressive fusion network) [31].** This model proposes a progressive fusion network based on semantic information to detect fake news. It first captures representative information from different modalities at different levels and achieves modality fusion within and across levels through a mixer, thus establishing strong connections between modalities.

• **BDANN (BERT-based domain adaptation neural network) [32].** The model utilizes a pre-trained BERT model to extract textual features, pre-trained VGG-19 model to extract visual features, and incorporates a domain classifier to mitigate dependency on specific events.

• **Roberta+CNN [33].** This framework integrates specialized convolutional neural network models for image analysis and sentence transformers for text analysis. By embedding features extracted from visual and textual inputs into dense layers, it ultimately enables accurate prediction of deceptive imagery.

• **FSRU (Frequency spectrum representation and fusion network) [34].** An architecturally simple and computationally efficient multimodal spectrum rumor detector is proposed, incorporating modules such as Fourier Transform and Spectrum Consistency, which enhances multimodal learning by introducing dual contrastive learning, combining text and image embedding, multimodal frequency spectrum representation and fusion module, and generating a cohesive multimodal representation based on feature distribution similarity.

4.5 Performance Comparison

We employ accuracy, precision, recall, and F1 score to evaluate the performance of the models. The experimental comparison results between the proposed model and different baselines are shown in Table 4 and Fig. 5.

Table 4: The comparison of experimental results

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	Textual	0.644	0.670	0.609	0.638	0.621	0.681	0.649
	Visual	0.599	0.608	0.621	0.614	0.588	0.575	0.581
	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
	MPFN	0.833	0.846	0.921	0.880	0.809	0.721	0.740
	BDANN	0.830	0.810	0.630	0.710	0.830	0.930	0.880
	Robert+CNN	0.853	0.821	0.943	0.877	**0.913**	0.745	0.820
	FSRU	**0.952**	**0.983**	0.938	**0.960**	0.901	**0.984**	**0.940**
	MAGF	0.918	0.922	**0.972**	0.946	0.904	0.863	0.883
Weibo	Textual	0.724	0.720	0.759	0.739	0.728	0.686	0.707
	Visual	0.608	0.580	0.863	0.694	0.698	0.737	0.654
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MPFN	0.838	0.857	0.894	0.889	0.873	0.863	0.876
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	Robert+CNN	0.812	0.851	0.784	0.816	0.744	0.826	0.782
	FSRU	**0.901**	**0.922**	0.892	**0.906**	0.879	**0.913**	**0.895**
	MAGF	0.887	0.859	**0.935**	0.895	**0.924**	0.837	0.878

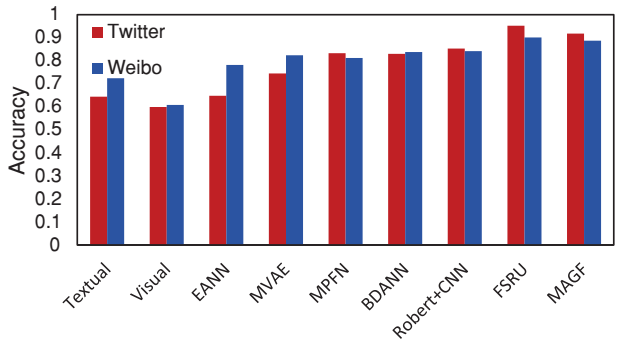


Figure 5: Comparison of results from different methods on two datasets

In the detection of fake news in a single modality, the text modality is significantly better than the image modality. This is because text can more accurately express the core content of the event, and the semantic information within it directly helps to identify fake news. In contrast, the visual information provided by images is relatively limited and cannot provide as rich semantics as text. Therefore, text has more advantages in fake news detection. Although single-modal models show some effectiveness in fake news detection, they still have certain deficiencies compared with multi-modal models. Multi-modal fake news detection methods utilize various information sources such as text and images to obtain more comprehensive and rich feature representations.

In the realm of multi-modal methods, MVAE achieves a higher accuracy than EANN, as it introduces encoders and decoders to learn shared representations of text and image, and better understand the complexity of the data. MPFN exhibits even higher performance, indicating that the attention mechanism can help improve the model's performance by considering the image parts relevant to the text. BDANN incorporates a domain classifier, exhibiting strong performance in learning and leveraging shared aspects between events, thus demonstrating robust capabilities in fake news detection tasks. Robert+CNN employs a novel method to image analysis and uses sentence transformers on text. However, both BDANN and Robert+CNN overlook the deep semantic relationships and interactions between features, leading to limitations in their detection accuracy. FRSU, by extracting rumor evidence from both unimodal and cross-modalperspectives hidden within frequency components, captures distinctive unimodal features and integrates cross-modal evidence of rumor accuracy in the frequency domain, demonstrating exceptional performance.

The proposed MAGF achieves significant advantages after introducing vocabulary and inter-region message passing and cross-modal gated fusion networks. The accuracy on two datasets is improved to 91% and 89%, respectively, validating the effectiveness of the proposed approach. On the Twitter dataset, the recall for fake news detection surpasses the latest research progress by 3.4%. On the Weibo dataset, the recall and precision for fake news detection exceed the latest research progress by 4.3% and 4.5%, respectively. To avoid generating redundant data, MAGF integrates cross-modal messages, transferring the most significant features between text and visual modalities. During the fusion process, text features and image region features that have undergone message passing are input into the cross-modal gated fusion network. By introducing gating units to adaptively control cross-modal feature fusion, MAGF achieves precise feature fusion and avoids interference from mismatched information, thereby enhancing the accuracy of fake news detection.

4.6 Ablation

4.6.1 Quantitative Analysis

We designed three model variants for three sets of experiments. The three variants of MAGF are the model without cross-modal message aggregation (MAGF w/o A), the model without gated fusion (MAGF w/o G), and the model without collaborative attention (MAGF w/o C). Through these three sets of experiments, we can analyze and compare the contributions of each component in our model, further understanding the role of each component in fake news detection.

(1) MAGF w/o A. BERT is used to extract word-level and contextualized text features. For images, Faster R-CNN is first employed to generate candidate region objects, and ResNet is then applied to extract regional features from these objects. Meanwhile, VGG-19 is used to extract global features from the entire image. The word-level text features and image regional features are processed through a gated fusion network, and fused with the image global features using

a collaborative attention mechanism to obtain a multi-modal feature representation, which is fed into the fake news detector as input.

(2) MAGF w/o G: Similarly, BERT is utilized to extract word-level and contextualized text features. For images, Faster R-CNN generates candidate region objects, and ResNet extracts regional features from them, while VGG-19 extracts global features from the whole image. Afterward, the word-level text features and image regional features are aggregated to combine information between regions and words. This aggregated feature is then fused with the image global features using the collaborative attention mechanism to obtain a multi-modal feature representation, which serves as input for the fake news detector.

(3) MAGF w/o C. BERT is again used to extract word-level and contextualized text features. For images, Faster R-CNN generates candidate region objects, and ResNet extracts regional features from these objects. The word-level text features and image regional features undergo message aggregation, and the resulting features are then fed into a gated fusion network to obtain fused text and image features. These fused features are simply concatenated to form a feature vector, which is used as input for the fake news detector.

The comparison of results between MAGF and the three model variants on the two datasets is shown in Table 5.

Table 5: Comparison of result of ablation experiments

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	MAGF w/o A	0.866	0.853	0.873	0.863	0.878	0.859	0.868
	MAGF w/o G	0.855	0.833	0.899	0.865	0.883	0.809	0.844
	MAGF w/o C	0.882	0.870	0.906	0.887	0.895	0.856	0.875
	MAGF	**0.918**	**0.922**	**0.971**	**0.946**	**0.904**	**0.863**	**0.875**
Weibo	MAGF w/o A	0.847	0.831	0.882	0.856	0.866	0.810	0.837
	MAGF w/o G	0.843	0.846	0.850	0.848	0.839	0.835	0.837
	MAGF w/o C	0.866	0.852	0.897	0.874	0.884	0.834	0.858
	MAGF	**0.887**	**0.859**	**0.935**	**0.895**	**0.924**	**0.837**	**0.878**

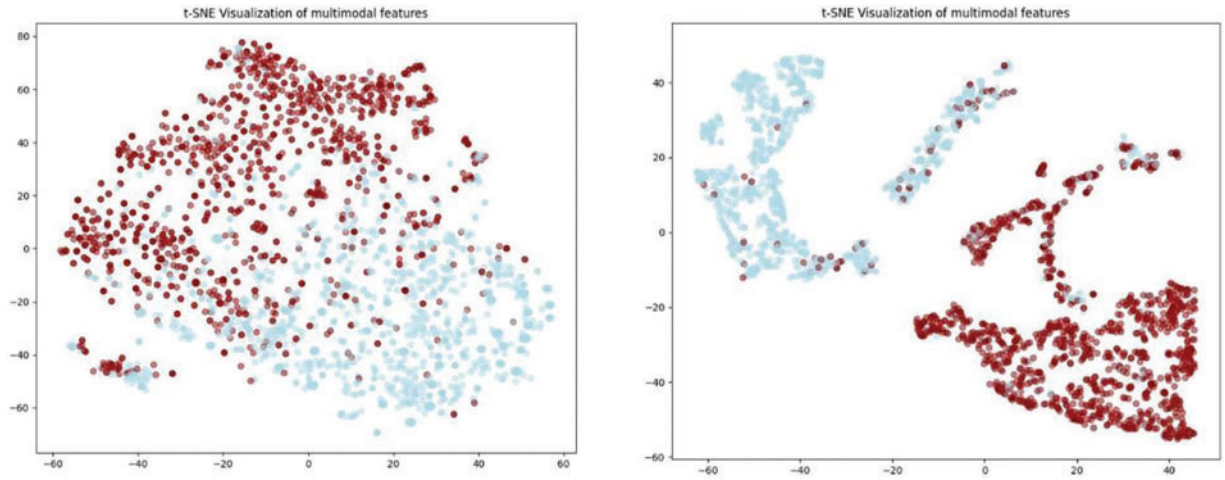
From the results presented in Table 5, we can make the following observations: 1) The MAGF w/o A variant achieved poorer performance, highlighting the importance of cross-modal message aggregation in fake news detection. This module effectively aggregates information between regions and words and reinforces key messages that need to be communicated across different modalities. This ensures that important information is fully considered and utilized in subsequent steps of the model, thus improving its performance. 2) The MAGF w/o G variant also exhibited lower performance, indicating that the gated fusion module significantly contributes to the model's performance. By dynamically adjusting the gate values, this module suppresses the fusion of image region features that do not match the text, allowing for a more effective integration of relevant information between text and images. This effectively reduces interference from irrelevant information. 3) Compared to MAGF, the MAGF w/o C variant showed relatively poor performance, suggesting that the collaborative attention fusion module, which learns multi-modal features, is more effective compared to the approach of simply concatenating features to create multi-modal representations.

These observations demonstrate that the complementary and synergistic effects of the various modules are crucial for the effectiveness of fake news detection. By comprehensively utilizing textual features, global image features, and regional image features, combined with optimization methods such as cross-modal message aggregation, gated fusion networks, and collaborative attention fusion, the model is able to achieve outstanding performance in fake news detection on both datasets.

4.6.2 Qualitative Analysis

To further analyze the effectiveness of the gated fusion network in MAGF, we used T-SNE [35] to qualitatively visualize the multi-modal features learned by MAGF w/o G and MAGF on the Weibo test set, as shown in Fig. 4.

From Fig. 6, it becomes evident that the separability of the multi-modal feature representations learned by MAGF is superior to MAGF w/o G. As depicted in Fig. 6a, while MAGF w/o G is able to learn distinguishable features, many of them are still prone to misclassification. In contrast, MAGF utilizes multi-modal information more precisely through gated fusion, resulting in a more distinct separability of the learned feature representations in Fig. 6b. Additionally, there are larger isolated regions in Fig. 6b. Therefore, the introduction of gated fusion and its integration with other components enable our model to excel in distinguishing between real and fake news.



(a) MAGF w/o G (b) MAGF

Figure 6: Visualization of multi-modal features

4.7 Case Study

The objective of Content-Based Fake News Detection (CBFND) is to assess the intent of news, as a set of quantifiable features extracted from news content, typically machine learning features. CBFND is a crucial tool for identifying the harmfulness of news [36]. Both text features and image features can be part of CBFND. In order to emphasize the significance of multimodal features in fake news detection, we conducted a comparative analysis between the results obtained from MAGF and those from unimodal models. Figs. 7 and 8 showcase fake news instances correctly identified by MAGF but overlooked by unimodal models.

New species of fish found at Arkansas.

Sharks in the streets of New York and New Jersey because of hurricane Sandy.

Figure 7: Fake news detected correctly by MAGF but incorrectly by the text

This is a species of water mantis with a highly toxic head, two to three times more potent than that of a pufferfish. There is no cure once poisoned.

Mr. Run Run Shaw has passed away. We have thousands of school buildings named "Run Run Shaw Teaching Building," and many hospitals have "Run Run Shaw Outpatient Building." Throughout his lifetime, this gentleman donated billions upon billions of medical and educational facilities.

Figure 8: Fake news detected correctly by MAGF but incorrectly by the image

Fig. 7 displays two high-confidence tweets successfully detected by MAGF but missed by pure text. The textual evidence in these two examples is sparse, making it difficult for the model to classify them as fake news. However, the accompanying images appear to be manipulated and inconsistent with the textual content.

In Fig. 8, two examples were successfully detected by MAGF but missed by the pure image model. The accompanying images appear normal; however, the textual content is exaggerated and illogical. This situation poses a challenge for MAGF based solely on images, but our MAGF model, leveraging multimodal features, is able to classify them correctly.

From these comparative cases, we can draw a conclusion that when the singlemodal model fails to distinguish false news, the detection accuracy can be improved by using MAGF based on multi-modal features.

4.8 Analysis and Discussion

In our study, the adaptive message passing and the gated fusion network played a crucial role in enhancing the model's performance. By integrating textual and visual information, our approach

significantly improved the accuracy and robustness of fake news detection. Specifically, the cross-modal message aggregation module enabled the model to recognize complex relationships between text and images, while the gated fusion network adaptively controlled the information flow, ensuring that only relevant information was retained for decision-making. Our research demonstrates innovation in several aspects. Firstly, unlike traditional methods, our modal not only considers the global features of images and text but also emphasizes the extraction of local features. The experimental findings demonstrate the significance of extracting both global and local features from images in the context of fake news detection. Global features provide the overall context of the image, while local features help detect details that are inconsistent with the textual description. This method effectively identifies contradictions between images and text. Secondly, the cross-modal message aggregation module achieves deeper information interaction between text and images. Finally, the gate-controlled fusion network, through its adaptive mechanism, filters out irrelevant information, allowing for the full utilization of cross-modal data, which further enhances the detection performance and robustness of the model. These innovations collectively strengthen the model's performance in the task of fake news detection.

Despite the significant effectiveness of our model, there are still some limitations. Firstly, relying on pooling operations to extract global features may lead to the loss of some important image information. Specifically, pooling operations compress image data to reduce computational load, but in the process, some details and local information may be overlooked. These details are crucial for detecting subtle inconsistencies between images and text. For example, if a small region in the image contains important information directly related to the textual description, pooling operations might discard it, preventing the model from effectively capturing this critical information. To mitigate this issue, we also extracted local features of the images to retain more detailed information, but we cannot guarantee that all important information is preserved. Therefore, future research could explore other methods to ensure that significant details are not overlooked during feature extraction. Secondly, the generalization ability of the model needs further verification when handling different types of data. Our experiments were mainly based on specific datasets, and the applicability and robustness of the model need to be validated for news data from different sources and types. This also provides a direction for future research, which can involve experiments on more diverse datasets to verify the model's generalizability and stability.

5 Conclusion

The interaction between cross-modal information and the effective utilization of key information play a critical role in multimodal fake news detection. We first formulate the task of cross-modal information aggregation. Based on this, we utilize the aggregated information to calculate gate values for dynamic fusion, focusing particularly on the effective information between text and images. A multimodal fake news detection model is constructed based on cross-modal message aggregation and gated fusion. Through experimental comparisons of different features, the results show that utilizing the deep interaction between text features and image regional features, as well as global features, can effectively distinguish between fake news and real news. Extensive experiments have been conducted on publicly available Weibo and Twitter datasets, verifying the effectiveness of our model. However, we have only explored text and image features at this stage. In the future, we will further explore how to incorporate other informational features in news into our consideration and seek better methods to enhance the extraction of detailed information, to further enhance the model's performance and applicability, certain measures can be taken.

Acknowledgement: The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (No. 62302540), with author Fangfang Shan. For more information, please visit their website at https://www.nsfc.gov.cn/ (accessed on 31/05/2024). Additionally, it is also funded by the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022020), where Fangfang Shan is an author. Further details can be found at http://xt.hnkjt.gov.cn/data/pingtai/ (accessed on 31/05/2024). The research is also supported by the Natural Science Foundation ofHenan Province Youth Science Fund Project (No. 232300420422), and for more information, you can visit https://kjt.henan.gov.cn/2022/09-02/2599082.html (accessed on 31/05/2024).

Author Contributions: Study design: Fangfang Shan, Mengyao Liu; Data analysis: Mengyao Liu; Data collection: Menghan Zhang, Zhenyu Wang; Analysis of experimental results: Mengyao Liu; Manuscript writing: Mengyao Liu; Manuscript guidance and revision: Fangfang Shan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Twitter data used to support the findings of this study have been deposited on the website: http://www.multimediaeval.org/mediaeval2016/verifyingmultimediause/index.html (accessed on 31/05/2024). The Weibo data used to support the findings of this study have been deposited on the website: https://github.com/wangzhuang1911/Weibo-dataset (accessed on 31/05/2024).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Z. Zhang, J. Cao, and D. Radcliffe, "The impact of COVID-19 on journalism in emerging economies and the global south," *Youth J.*, vol. 699, no. 7, pp. 99–100, 2021.

[2] I. Rish, "An empirical study of the naive Bayes classifier," presented at the IJCAI 2001, Washington, USA, Aug. 4–6, 2021, vol. 3, pp. 41–46. doi: 10.1002/9781118721957.ch4.

[3] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991. doi: 10.1109/21.97458.

[4] W. S. Noble, "What is a support vector machine?" *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006. doi: 10.1038/nbt1206-1565.

[5] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2915–2921. doi: 10.1145/3308558.3313552.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017. doi: 10.1145/3137597.3137600.

[7] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: A transformer-based approach," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 335–362, 2022. doi: 10.1007/s41060-021-00302-z.

[8] Y. Wang, L. Wang, Y. Yang, and T. Lian, "SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection," *Expert. Syst. Appl.*, vol. 166, no. 166, pp. 114090, 2021. doi: 10.1016/j.eswa.2020.114090.

[9] S. I. Manzoor, J. Singla, and Nikita, "Fake news detection using machine learning approaches: A systematic review," in *2019 3rd Int. Conf. Trends Electronics Inform. (ICOEI)*, Tirunelveli, India, 2019, pp. 230–234. doi: 10.1109/ICOEI.2019.8862770.

[10] Y. F. Huang and P. H. Chen, "Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms," *Expert. Syst. Appl.*, vol. 159, pp. 113584, 2020. doi: 10.1016/j.eswa.2020.113584.

[11] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 881–893, 2021. doi: 10.1109/TCSS.2021.3068519.

[12] X. C. Liu, "Research on fake news detection based on machine learning algorithms," *Inf. Technol. Inf.*, no. 9, pp. 237–239, 2021.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[14] T. Chen, X. Li, H. Yin, and J. Zhang,"Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Trends Appl. Knowl. Discov. Data Min.*, Melbourne, VIC, Australia, Jun. 3, 2018, pp. 40–52. doi: 10.1007/978-3-030-04503-6_4.

[15] J. Ma, W. Gao, and K. F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning,"presented at The World Wide Web Conf., San Francisco, USA, May 13–17, 2019, pp. 3049–3055. doi: 10.1145/3308558.3313741.

[16] T. Mahmood, Z. Mehmood, M. Shah, and T. Saba, "A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform," *J. Vis. Commun. Image Rep.*, vol. 53, no. 5, pp. 202–214, 2018. doi: 10.1016/j.jvcir.2018.03.015.

[17] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," presented at the 2019 IEEE ICDM, Beijing, China, Nov. 8–11, 2019, pp. 8–11. doi: 10.1109/ICDM.2019.00062.

[18] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo and J. Li, "Exploring the role of visual content in fake news detection," *Disinf. Misinf. Fake News Social Media: Emerg. Res. Chall. Oppor.*, pp. 141–161, 2020. doi: 10.1007/978-3-030-42699-6.

[19] Z. W. Jin, J. Cao, H. Guo, and Y. D. Zhang, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," presented at the Pro. 25th ACM Int. Conf. Multimed., Paris, France, Oct. 23, 2017, pp. 795–816. doi: 10.1145/3123266.3123454.

[20] Y. Q. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," presented at the Proc. 24th ACM SIGKDD Int. Conf. Data Min Knowl. Discov, London, UK, Jul. 19, 2018, pp. 849–857. doi: 10.1145/3219819.3219903.

[21] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh, "SpotFake: A multi-modal framework for fake news detection," presented at the 2019 IEEE Fifth Int. Conf. Multimed. Big Data, Changsha, China, Sep. 11–13, 2019, pp. 39–47. doi: 10.1109/BigMM.2019.00-44.

[22] S. S. Qian, J. G. Wang, and J. Hu, "Hierarchical multi-modal contextual attention network for fake news detection," presented at the Proc. 44th Int. ACM SIGIR Conf., Canada, Jul. 11–15, 2021, pp. 153–162. doi: 10.1145/3404835.3462871.

[23] Y. Wu, P. W. Zhan, Y. J. Zhang, L. M. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," presented at the ACL-IJCNLP 2021, Bangkok, Thailand, Aug. 1–8, 2021, pp. 2560–2569. doi: 10.18653/v1/2021.findings-acl.226.

[24] F. F. Shan, H. F. Sun, and M. Y. Wang, "Multimodal social media fake news detection based on similarity inference and adversarial networks," *Comput. Mater. Contin.*, vol. 79, no. 1, pp. 581–605, 2024. doi: 10.32604/cmc.2024.046202.

[25] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016. doi: 10.1109/TPAMI.2016.2577031.

[26] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition,"presented at the 2016 IEEE CVPR, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Adv. Neural Inf. Process. Syst.*, NY, USA, 2019, vol. 32. doi 10.48550/arXiv.1908.02265.

[28] C. Maigrot, V. Claveau, E. Kijak, and R. Sicre, "MediaEval 2016: A multimodal system for the verifying multimedia use task," in *MediaEval 2016: "Verfiying Multimedia. Use" Task*, Hilversum, Netherlands, Oct. 20–21, 2016.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[30] J. T. Zhuang *et al.*, "AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18795–18806, 2020. doi: 10.48550/arXiv.2010.07468.

[31] J. Jing, H. C. Wu, J. Sun, X. C. Fang, and H. X. Zhang, "Multimodal fake news detection via progressive fusion networks," *Inf. Process. Manage.*, vol. 60, no. 1, pp. 103120, 2023. doi: 10.1016/j.ipm.2022.103120.

[32] T. Zhang *et al.*, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection,"presented at the 2020 Int. Joint Conf. Neural Netw, Glasgow, UK, Jul. 19–24, 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206973.

[33] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, 2022. doi: 10.1007/s00521-021-06086-4.

[34] L. An *et al.*, "Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector," presented at the AAAI Conf. Artif. Intell., Feb. 22–25, 2024, vol. 38, pp. 18426–18434. doi: 10.1609/aaai.v38i16.29803.

[35] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[36] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, "Content-based fake news detection with machine and deep learning: A systematic review," *Neurocomputing*, vol. 530, pp. 91–103, 2023. doi: 10.1016/j.neucom.2023.02.005.