frontiers in Physics

ORIGINAL RESEARCH
published: 24 September 2021
doi: 10.3389/fphy.2021.711221

Multilevel Attention Residual Neural Network for Multimodal Online Social Network Rumor Detection

*Zhuang Wang and Jie Sui**

School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China

In recent years, with the rapid rise of social networks, such as Weibo and Twitter, multimodal social network rumors have also spread. Unlike traditional unimodal rumor detection, the main difficulty of multimodal rumor detection is in avoiding the generation of noise information while using the complementarity of different modal features. In this article, we propose a multimodal online social network rumor detection model based on the multilevel attention residual neural network (MARN). First, the features of text and image are extracted by Bert and ResNet-18, respectively, and the cross-attention residual mechanism is used to enhance the representation of images with a text vector. Second, the enhanced image vector and text vector are concatenated and fused by the self-attention residual mechanism. Finally, the fused image–text vectors are classified into two categories. Among them, the attention mechanism can effectively enhance the image representation and further improve the fusion effect between the image and the text, while the residual mechanism retains the unique attributes of each original modal feature while using different modal features. To assess the performance of the MARN model, we conduct experiments on the Weibo dataset, and the results show that the MARN model outperforms the state-of-the-art models in terms of accuracy and F1 value.

Keywords: online social networks, rumor detection, neural networks, multimodal fusion, attention residual network

OPEN ACCESS

Edited by:
Shudong Li,
Guangzhou University, China

Reviewed by:
Keke Huang,
Central South University, China
Chengyi Xia,
Tianjin University of Technology, China
Yan Wang,
Macquarie University, Australia

***Correspondence:**
Jie Sui
suijie@ucas.ac.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *18 May 2021*
Accepted: *03 August 2021*
Published: *24 September 2021*

Citation:
Wang Z and Sui J (2021) Multilevel Attention Residual Neural Network for Multimodal Online Social Network Rumor Detection. Front. Phys. 9:711221. doi: 10.3389/fphy.2021.711221

INTRODUCTION

Since the beginning of the 21st century, with the rapid development of the Internet technology and the gradual popularization of computers and other network terminal equipment, the dissemination speed of all kinds of news has been a qualitative leap, which has changed the inherent living habits of human beings to a certain extent. Especially after 2004, with the advent of the Web2.0 era [1], online social media represented by Facebook, Twitter, and Sina Weibo have developed rapidly, which not only have a great impact on the traditional news industry but also facilitate people's access to news.

Compared with the traditional news industry, social media have a lower release threshold, a faster spread, and a wider range of influence. These network rumors reduce the quality of people's access to information and seriously endanger the security of the whole society and even at the national level. In particular, rumors about some major public emergencies can easily cause panic and social unrest. Take the COVID-19 transmission incident in early 2020 as an example, from "COVID-19 is the evolutionary version of SARS" to "double coptis can prevent coronavirus infection;" rumors about the event emerge endlessly, which greatly hinder the overall prevention and control of the epidemic and causes adverse social effects.

Recently, many social media have been allowing users to add corresponding images or videos while publishing texts. News with images is more confusing and disseminating, and its forwarding frequency is 11 times more than that of pure text news [2]. However, most of the existing rumor detection models only focus on the propagation path or text of the news but ignore the images related to the event. At present, only a few works focus on the image in the news, but generally these multimodal rumor detection models only simply concatenate image features and text features for classification. In fact, the semantic features of each mode are heterogeneous in the feature space, which may lead to the following two problems:

1) The fusion of multimodal features is insufficient.
2) The noise information generated by the fusion is large, which affects the final classification results.

To solve these problems, we propose a multimodal social network rumor detection model based on the multilevel attention residual neural network. Among them, the multilevel attention mechanism selectively fuses the image–text features from the semantic level. Compared with the traditional fusion method of image–text features, the proposed method greatly improves the joint representation performance between different modal features. The residual structure retains the unique attributes of different modal features on the basis of the image–text joint representation, which effectively alleviates the noise information caused by different modal fusions. The contribution of this article can be summarized as the following three points:

1) This article proposes a multimodal social network rumor detection model based on the multilevel attention residual neural network.
2) The multi-layer attention mechanism improves the feature fusion effect between multiple modalities, and the residual structure effectively alleviates the adverse effects of the noise information generated during the fusion.
3) The experimental results on the real Weibo dataset show that the accuracy and F1 value of the MARN model are higher than those of the current mainstream multimodal rumor detection models.

RELATED WORK

Concept and Development of Rumor

The spread of rumors is a social phenomenon that develops with the development of the human society. It is often used as a weapon by hostile parties to fight. It has long been a hot topic of research. The systematic research on rumors began with Alport and Postman's [3] *The Psychology of Rumors*, which defines rumors as statements of information on specific or current topics that tend to spread from person to person, usually by oral means, without any evidence to prove their authenticity.

Compared with traditional rumors, network rumors have some different characteristics, such as faster spread and wider impact, which also brings great challenges to the detection of network rumors. The most original measures to prevent and control network rumors are by basically using a combination of user reports and manual verification for rumor detection and tracking, which not only consumes a large amount of human resources but also has a strong time lag. It is often difficult to predict and eliminate the rumors in the early stage of spread. To solve these problems, in recent years, a large number of scholars have used machine learning or neural network learning methods to detect rumors on Weibo and Twitter news and have achieved a series of results.

Research Status at Home and Abroad

From the data sources of the model, rumor detection can be roughly divided into two categories: propagation-based and content-based rumor detection. The former is based on the principle of the network structure [4–7] and uses the propagation path of the posts to classify them [8]. The latter is to use the post or its additional modal information for classification. The rumor detection referred to in this article is all content-based. It can generally be divided into three types: traditional machine learning–based methods, unimodal feature–based neural network methods, and multimodal feature–based neural network methods.

Rumor Detection Based on Traditional Machine Learning

Castillo et al. [9], who first introduced the machine learning method to the field of network rumor detection, used a variety of traditional machine learning methods to detect the reliability of the datasets collected on Twitter and achieved some results. The first one to automatically detect rumors on Sina Weibo was the method suggested by Yang et al. [10], which uses the SVM (Support Vector Machine, SVM) classifier to test and classify the datasets collected from Sina Weibo's official rumor development platform and proposes new detection features for the differences between Chinese and English language characteristics, pioneering the rumor detection in Chinese social networking platforms. On the basis of the above two studies, many experts and scholars [11, 12] have added text features, user features, and propagation features for rumor detection, which all improve the performance of rumor detection to a certain extent. Rumor detection based on traditional machine learning pioneers automated rumor detection and has a profound impact on the technological development of this field. However, such methods also have some drawbacks, such as the selection of indicators depends heavily on the experimenter's experience and the accuracy of the model's classification needs to be improved. This is also an important problem to be solved based on the neural network model.

Rumor Detection of the Neural Network Based on a Unimodal Feature

With the continuous progress of neural network technology in the field of natural language processing [13], more and more scholars have applied it to the field of rumor detection [14]. Ma et al. [15] applied the RNN (Recurrent Neural Network) model to

network rumor detection for the first time, which greatly improves the efficiency of rumor detection compared with traditional machine learning methods. Liu et al. [16] proposed an improved CNN (convolutional neural network) model for microblog rumor detection. The model is simple and easy to implement. Chen et al. [17] combined the attention mechanism with the RNN model for rumor detection, which solved the problem of excessive redundancy of text features and weak remote information connection to some extent. In 2019, Chen et al. [18] proposed an attention residual neural network combined with the CNN network for social network rumor detection, which is the first model to combine an attention model with a residual network for social network rumor detection. Experiments on two Twitter datasets show that the attention residual network can capture long-term dependencies and achieve high classification accuracy and F1 value regardless of the choice of policy. However, these traditional neural network models only focus on the text feature of rumors, ignoring the accompanying images and social characteristics, which limits the detection performance of the model and needs to be improved to adapt to the rapid development of the network era.

Rumor Detection of the Neural Network Based on Multimodal Features

Similar to sentiment classification [19, 20], social network rumor detection tasks have also entered the multimodal era in recent years. In 2017, Jin et al. [21] first introduced image features into fake news detection and created a corresponding multimodal microblog rumor dataset. The model first extracts event-related image semantic features through a pre-trained VGG19 (Visual Geometry Group, VGG) model and uses an attention mechanism to extract key features in the text and social context and then multiplies them element by element with image semantic features to adjust the weight of the visual semantic features. Experiments show that this method can detect many fake news cases which are difficult to distinguish under a unimodal feature. Wang [22] proposed an event-based antagonism network based on the work of the former. The multimodal feature extractor in this network is forced to learn the invariant representation of events to deceive the discriminator. In this way, it eliminates the strong dependency on specific events in the collected datasets and gains better generalization capabilities for unknown events. Dhruv et al. [23] then constrained the fused multimodal vectors through an automatic encoder to better learn the joint representation. Liu et al. [24] made full use of the text information contained in the image and improved the detection performance of the model by extracting hidden texts from the image.

Summarizing the previous research, it can be found that these multimodal rumor detection models using image and text features have become a major trend in the field of rumor detection. Compared with the traditional pure text rumor detection models, the multimodal rumor detection models can effectively make use of the feature differences between different modes to complement each other and improve the performance of rumor detection. However, due to the huge semantic gap and redundant information among the modal features, the existing models still have the problem of insufficient feature fusion among the modes and huge noise information when fusing, which is also an important problem to be solved by our model.

PROBLEM STATEMENT

In essence, the detection of rumors in social media is a two-category problem. That is, the experimenter divides the input content into rumors or non-rumors through a specific model. If the input content is a series of information such as the post itself and its related comments, forwarding, etc., it is called event-level rumor detection; if the input content is just the post, it is called post-level rumor detection. For example, user U1 posts a post, user U2 comments on the post, and user U3 retweets the post. Event-level rumor detection uses all of this relevant information as the basis for rumor detection, while post-level rumor detection uses only the posts posted by user U1. Our model belongs to post-level rumor detection, with the aim of identifying rumors in their early stages to avoid greater social harm.

We define a post $X = \{T, P\}$ as a tuple representing two different content patterns. $T = \{w_1, w_2, \ldots, w_n\}$ represents the text content contained in the post, where n is the number of words (w). $P = \{p_1, p_2, \ldots, p_m\}$ represents the image content attached to the post, where m is the number of images (p). The true tag of a post is $y = \{0, 1\}$ when $y = 0$ it means that the content of the post is true and when $y = 1$ the post is a rumor. Formally, rumor detection on the post-level aims to learn a projection $F(X) \rightarrow \{0, 1\}$

MODEL

In this section, cross-attention and self-attention mechanisms are used to enhance the fusion of image and text representation, and the residual mechanism is used to alleviate the adverse effects of the noise information generated during the fusion. We first describe the general framework of the model and then describe in detail the principle and operation of each component that makes up the model.

Building Model Framework

We propose a multimodal online social network rumor detection model based on the multilevel attention residual mechanism. Its overall framework is shown in **Figure 1** and consists of the following four parts:

1) Image–text embedding: The pre-trained models ResNet-18 and Bert are used to extract the original features of images and texts and transform them into the corresponding vectors R_P and R_T.
2) Cross-attention residual module: The text vector R_T is used to enhance the representation of the image vector R_P by cross-attention, and then the residual mechanism is used to add the vector R_P to the enhanced picture vector to get the vector R'_P
3) Self-attention residual module: The concatenated image–text vector R_C is fused by the self-attention process and sliced, and

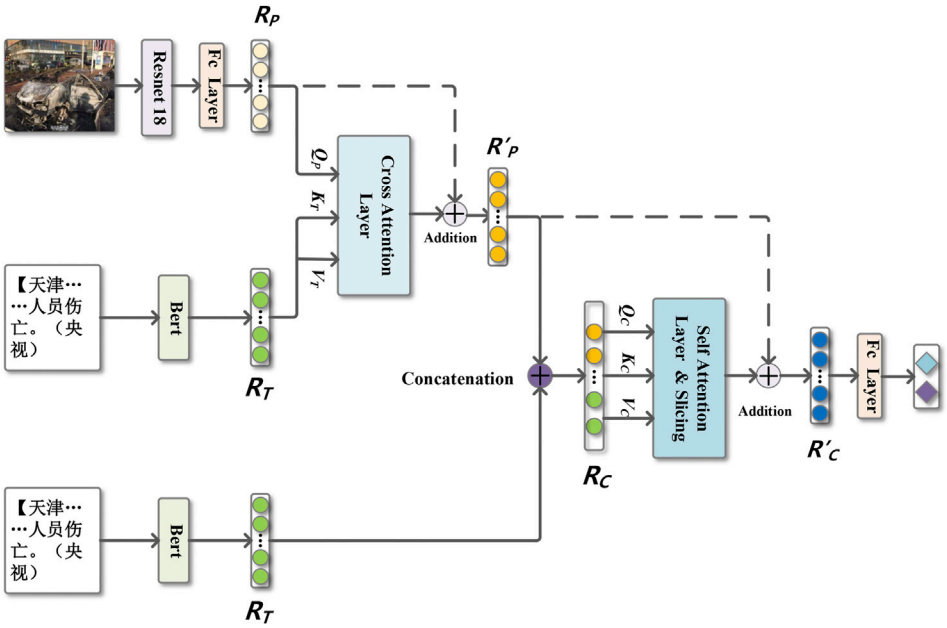

FIGURE 1 | Network structure of the MARN model. ***Instruction***: Dotted lines represent residual connections.

then the vector R'_P is added to the sliced vector to get the vector R'_C by using the residual mechanism.

4) Rumor classifier: It consists of a fully connected layer that binds the vector R'_C to get the final result.

Defining Image–Text Embedding

Using the Bert Text Extractor

Our model uses the pre-trained model Bert [25] as a feature extraction method, which improves the performance significantly compared with the traditional language model. The main reason is that Bert proposed a new pre-trained target and masked language model, which randomly masks 15% of the words in each sentence and uses the context to encode them in both directions, enriching the contextual feature of each word. In addition, Bert also pre-trains whether the two sentences are continuous. Specifically, Bert selects some sentences for A and B during the pre-training process, where statement B has a 50% probability of being the next sentence in statement A and a 50% probability of being randomly selected in the corpus. The goal is for Bert to learn the relevance of the two sentences and to better accommodate downstream tasks that require an understanding of the relationship between the upper and lower sentences.

The overall structure of the Bert model is shown in **Figure 2** [25], which is mainly composed of three parts: embedding layer, coding layer, and output layer. The embedding layer consists of three parts: token embedding, sentence embedding, and position embedding, which represent the word vector of the word, which sentence it belongs to, and where it is in the sentence. The coding layer is composed of the encode parts of a multilayer transformer. Through the multi-head attention mechanism and residual module, Bert can better enhance the extraction of deep semantic features of the text. There are two forms of the output layer where one is the vector *encode_out*, which represents the features of the whole statement. The other is the vector *pooled* representing the information of the first position [CLS]. In this model, the first output form is used. Each sentence can get a text vector $R_T \in \mathbb{R}^{\text{pad}\times d_T}$ after the Bert model, where represents the number of words embedded in each sentence and d_T represents the embedding dimension of each word.

Using ResNet-18 Image Extractor

This section uses a deep residual network ResNet-18 model based on transfer learning to extract image features. Compared with a traditional VGG model, the ResNet-18 model has smaller parameters, faster training speed, and higher accuracy. The ResNet model was originally proposed by Kaiming He et al. [26] and is widely used in image processing and computer vision. The main idea is to use multilevel residual modules to connect, which effectively alleviates the disappearance of back propagation gradient and model performance degradation caused by too many layers in traditional deep convolutional neural network models.

Each residual unit consists of a residual learning branch and an identical mapping branch, the structure of which is shown in **Figure 3** [26]. Here, i is the input, $G(i)$ is the result of the residual learning branch, ReLU is the activation function, and the output of the residual unit can be expressed as $H(i) = G(i) + i$. When the residual learning branch does not work, it can be expressed as $H(i) = i$. The two 1*1 layers in the residual branch function to reduce and increase the dimension of $G(i)$ to ensure that the dimension of $G(i)$ is consistent with that of i for subsequent operations.

We extract the last layer of the feature vector D in the ResNet-18 model by first stretching it through the flatten layer and then extracting the image vector R_P through a fully connected layer,

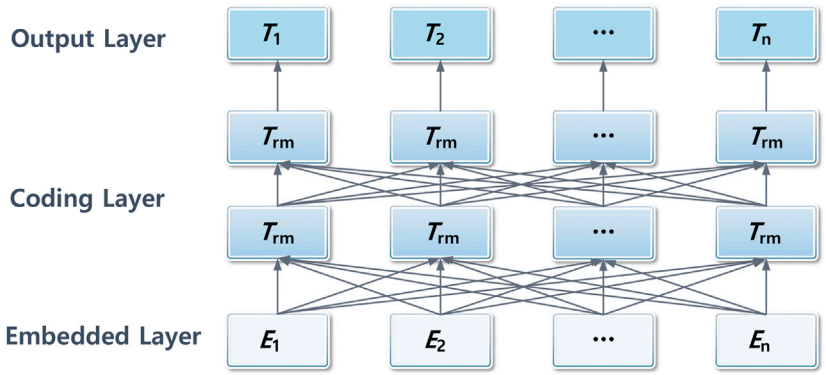

FIGURE 2 | Structure of the Bert model.

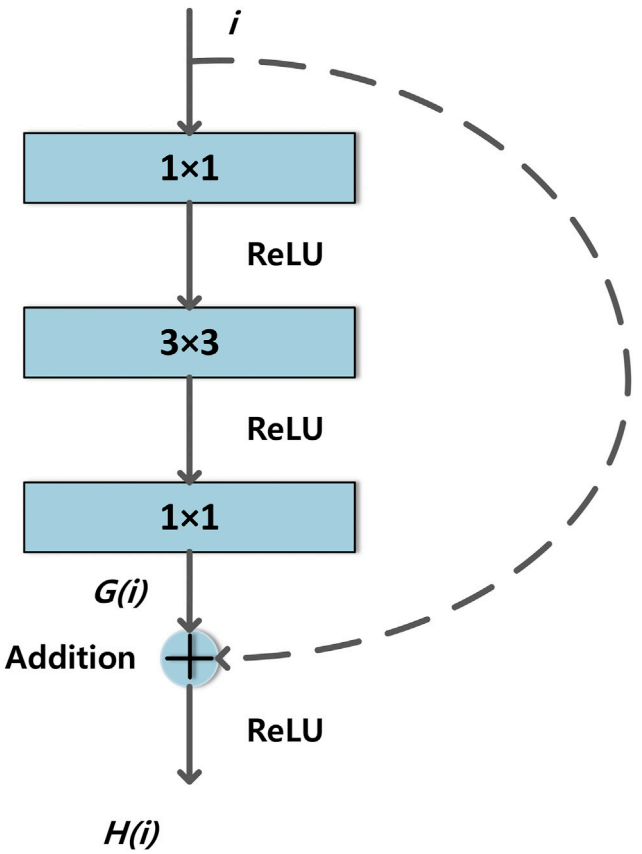

FIGURE 3 | Structure of the ResNet-18 unit. ***Instruction***: Dotted lines represent residual connections.

$$R_P = \mathrm{ReLU}\left(\mathrm{Flatten}\left(D\right) \times W_D + b_D\right) \tag{1}$$

Where, $R_P \in \mathbb{R}^{m \times d_P}$, $D \in \mathbb{R}^{m \times d_D \times 1 \times 1}$, $W_D \in \mathbb{R}^{d_D \times d_P}$, W_D and b_D are the weight matrix and bias term of this fully connected layer, respectively, and the dimension of b_D is the same as that of R_P. ReLU is the activation function, and the function of the flatten layer is to stretch the multidimensional vector into one dimension. m is the number of images attached to each post, and the value taken is m = 1. d_D is the dimension of the last output vector of the ResNet-18 model, d_P is the output dimension after image extraction, and $d_P = d_T$.

Building A Cross-Attention Residual Module

Just like human vision, the attention mechanism [27] automatically gives greater weight to the more noteworthy parts. There are two reasons why the MARN model can enhance the fusion between images and texts. On the one hand, the powerful pre-training models can express the semantic level of common words or item shapes; on the other hand, the weight distribution of the attention mechanism itself can be continuously studied to obtain better results.

We use text vectors to enhance image vectors by the cross-attention mechanism.

First, define Q_P, K_T, and V_T as follows:

$$\begin{aligned} Q_P &= R_P \times W_{QP} \\ K_T &= R_T \times W_{KT} \\ V_T &= R_T \times W_{VT} \end{aligned} \tag{2}$$

Where $Q_P \in \mathbb{R}^{m \times d_K}$, $K_T \in \mathbb{R}^{pad \times d_K}$, $V_T \in \mathbb{R}^{pad \times d_V}$, $W_{QP} \in \mathbb{R}^{d_P \times d_K}$, $W_{KT} \in \mathbb{R}^{d_T \times d_K}$, $W_{VT} \in \mathbb{R}^{d_T \times d_V}$. d_K and d_V are the second dimensions of matrix W_{QP} (or W_{KT}) and W_{VT}, note that $d_K = d_V = d_P = d_T$.

Then, compute the enhanced image vector ATT_P:

$$ATT_P = \mathrm{Softmax}\left(\frac{Q_P \times K_T^{\top}}{\sqrt{d_K}}\right) \times V_T \tag{3}$$

Where $ATT_P \in \mathbb{R}^{m \times d_V}$, $K_T^{\top}$ is the transposition of vector K_T and Softmax is the normalization function.

Finally, in order to ensure that the performance of the image feature after attention enhancement is not inferior to that of the original vector R_P, the residual mechanism is used to fuse R_P with the enhanced image feature ATT_P to get the vector and accumulate the results of each fusion of m images. If the fusion effect between the image and text is not ideal, the model will automatically adjust the value of ATT_P through back propagation until it is very small so that R_P' is almost equal to R_P, ensuring that the effect after fusion will not deteriorate,

TABLE 1 | Statistics of the dataset.

	Rumor	Non-rumor	All
Training set	3,561	3,584	7,145
Testing set	1,187	1,195	2,382
All	4,748	4,779	9,527

$$R_P' = \sum_{\mathrm{m}} (ATT_P + R_P) \tag{4}$$

Where $R_P' \in \mathbb{R}^{1\times \mathrm{d}_V}$.

Compared with single-head attention, multi-head attention can learn the weight relationship between each element from different angles and then concatenate to get the final vector representation. Under normal circumstances, its performance is better than single-head attention. Our model uses multi-head attention for fusion, and the specific related parameters are shown in **Table 2**. Since its principle is the same as single-head attention, it will not be repeated here.

Building a Self-Attention Residual Module

After obtaining the image vector R_P' with enhanced text features, we will proceed to fuse the image–text features.

First, we concatenate the image vector R_P' with the text vector R_T to get the initial fusion vector R_C,

$$R_C = \mathrm{Concat}\left[R_P',\ R_T\right] \tag{5}$$

where $R_C \in \mathbb{R}^{(1+\mathrm{pad})\times \mathrm{d}_T}$ and Concat is the concatenation function.

Then, similar to the cross-attention residual module, define Q_C, K_C, and V_C,

$$\begin{aligned} Q_C &= R_C \times W_{QC} \\ K_C &= R_C \times W_{KC} \\ V_C &= R_C \times W_{VC} \end{aligned} \tag{6}$$

Where $Q_C \in \mathbb{R}^{(1+\mathrm{pad})\times \mathrm{d}_K}$, $K_C \in \mathbb{R}^{(1+\mathrm{pad})\times \mathrm{d}_K}$, $V_C \in \mathbb{R}^{(1+\mathrm{pad})\times \mathrm{d}_V}$, $W_{QC} \in \mathbb{R}^{\mathrm{d}_T\times \mathrm{d}_K}$, $W_{KC} \in \mathbb{R}^{\mathrm{d}_T\times \mathrm{d}_K}$, $W_{VC} \in \mathbb{R}^{\mathrm{d}_T\times \mathrm{d}_V}$.

Then, the self-attention mechanism is used to calculate the weight of the integrated vector R_C to obtain the enhanced integrated vector ATT_C,

$$ATT_C = \mathrm{Softmax}\left(\frac{Q_C \times K_C^{\top}}{\sqrt{\mathrm{d}_K}}\right) \times V_C \tag{7}$$

Where $ATT_C \in \mathbb{R}^{(1+\mathrm{pad})\times \mathrm{d}_V}$, $K_C^{\top}$ is the transposition of vector K_C.

Finally, use the residual mechanism to connect the vector ATT_C with the vector R_P'. It is worth noting that the dimension of ATT_C is not the same as that of vector R_P', so the vector ATT_C needs to be sliced before residual joining,

$$R_C' = \mathrm{Slice}\left(ATT_C\right) + R_C' \in \mathbb{R}^{\mathrm{d}_V} \tag{8}$$

Where $R_C' \in \mathbb{R}^{\mathrm{d}_V}$, Slice is the slicing function.

Defining Classifier and Loss Function

The vector R_C' is sent to the fully connected layer to obtain the prediction probability, $\hat{y}$

$$\hat{y} = \mathrm{Softmax}\left(R_C' \times W_C + b_C\right) \tag{9}$$

where $\hat{y}$ is the probability predicted by the model, W_C is the weight matrix of the fully connected layer, b_C is the bias term, $W_C \in \mathbb{R}^{\mathrm{d}_V\times 2}$ and the dimension of b_C is the same as $\hat{y}$.

We use cross-entropy as the loss function of this model, and the formula is as follows:

$$\mathrm{L}(\theta) = -\frac{1}{\mathrm{z}}\sum\nolimits_{X}\left[\mathrm{y}\ln\hat{y} + (1-\mathrm{y})\ln(1-\hat{y})\right] \tag{10}$$

where θ represents all parameters of the model, z is the total number of training samples, and y is the real label of samples.

We use the Adam optimizer to carry out back propagation, so as to obtain the best model parameters, and test the actual performance of the model on the testing dataset.

EXPERIMENTS

In this section, we first introduce the dataset and various hyper-parameters used in our experiment, then briefly introduce the baseline model we used, and finally analyze the results of the comparison experiment and the ablation experiment.

Dataset

In order to fairly compare the detection performance of this model and the baseline models, this article uses the Weibo dataset, which is commonly used in the field of multimodal rumor detection to carry out the experiment. This dataset was first published by Jin et al. [21], and it contains roughly the same number of rumor posts and non-rumor posts. Among them, rumor posts came from the official rumor debunking system of Weibo from May 2012 to January 2016 and non-rumor posts came from news verified by the authoritative Chinese news agency Xinhua News Agency. At the same time, in order to ensure the availability of the dataset, Jin et al. [21] deleted duplicated images and very small or very long images in the original image set. This article adopts the same method as paper [21], setting the ratio of the training set to test set to 4:1. The details of the dataset are shown in **Table 1**.

Hyperparameters

The experiment of this model is based on Python3.7, using the PyTorch deep learning framework, computing on GPU, and

TABLE 2 | Hyperparameters.

Hyperparameter	Value
Word embedding dimension (d_T)	768
Sentence length (pad)	64
Attention heads	96
Learning rate	0.001
Batch size	216
Dropout	0.2
Epochs	50

TABLE 3 | Results of different baseline models on a dataset.

Model	Accuracy	F1
Textual	0.8077	0.8074
Visual	0.6969	0.6954
att-RNN	0.7720	0.7685
MSRD	0.7940	0.7790
EANN	0.8270	0.8290
MVAE	0.8240	0.8230
MARN	**0.8581**	**0.8580**

Bold font represents the largest number in this column.

using cross-entropy loss function and Adam optimizer for back propagation optimization. At the same time, in order to prevent overfitting, a dropout layer is added after each fully connected layer to randomly delete some parameters when training the model. To save the training time and GPU memory space, we fixed the internal parameters of Bert and ResNet-18 models and did not participate in the back propagation training of the models. Other hyperparameters are shown in **Table 2**.

Baseline Models

To compare the performance of each model fairly, the following models are tested based on the above dataset, and the partition ratio of the training dataset and testing dataset is the same.

1) Textual Model

The textual model only uses the text features in the samples for experiments and directly transfers the text features into the Bert model for training, followed by two fully connected layers for classification.

2) Visual Model

The visual model only uses the image features in the sample for experiments and uses 0 to fill the sample with missing image features, that is, a pure black image is used to replace the image features in the sample. The image is encoded and input into the ResNet-18 model, followed by a dimension of 32 fully connected layers, and finally input into the classifier to get the sample classification results. In order to enhance the generalization ability of the model and reduce the training time, the ResNet-18 network adopts the method of migration learning, selects the model parameters that have been trained on the large dataset Image 1000, and does not participate in the back propagation. It only fine tunes the back wiring layer.

TABLE 4 | Results of different ablation models on a dataset.

Model	Accuracy	F1
MARN-CA-SA	0.8359	0.8357
MARN-CA	0.8472	0.8469
MARN-SA	0.8489	0.8486
MARN-residual	0.8484	0.8484
MARN	**0.8581**	**0.8580**

Bold font represents the largest number in this column.

3) Att-RNN

This model [21] uses the attention mechanism to fuse the text, image, and social features and then input them into the classifier for judgment. In order to make a fair comparison, we adopt the model after deleting the social characteristics, and the other parameters are consistent with those in the literature.

4) MSRD

In this model [24], First, the text in the image is extracted, and then it is connected with the text content in the sample. Finally, the image and the connected text are fused and classified at the feature level.

5) EANN

EANN [22] uses VGG19 and the Text-CNN (text-convolutional neural network) to extract the image and text features and uses the event discriminator to take the concatenated vector of constraints and finally input the concatenated vector to the classifier for classification.

6) MVAE

This model [23] uses the VAE (variational autoencoder) module to constrain the vector after multimodal feature fusion and then classify the feature vector.

7) MARN

The whole model is proposed in this article.

Comparison and Analysis of Baseline Models

We use common indicators such as F1 value and accuracy to evaluate each model. The results of each model are shown in **Table 3**.

Table 3 shows that the MARN model achieves 0.8581 and 0.8580 of the most important performance indicators F1 value and accuracy, respectively, which are higher than the mainstream multimodal rumor detection model and fully demonstrate the advanced performance of the MARN model. On the one hand, the multilevel attention mechanism selectively fuses the text and image features, making full use of the feature complementary function between each mode. On the other hand, the residual mechanism keeps the unique attributes of each mode while using the fused features, which ensures that the final result will not be worse than before.

In addition, it can be seen from **Table 3** that the accuracy and F1 value of the visual model are lower than those of the textual model. After all, in the current social network, text is still the most important source of information for people and images only play a minor role. Moreover, the performance of the textual model is better than that of the att-RNN model and MSRD model because we used Bert instead of the traditional LSTM (Long Short-Term Memory) as a text extractor. The EANN model and MVAE model use an event discriminator and a VAE, respectively, to constrain the concatenated vectors, which makes their accuracy higher. However, these two models only connect the vectors of different modes in series, so it is difficult to fuse the information of

different modes at the semantic level, which is what the MARN model focuses on.

Comparison and Analysis of Ablation Models

To further analyze the influence of each module on the overall model results, we deleted each module and carried out experiments.

MARN-CA-SA: The cross-attention residual module and self-attention residual module are deleted in this model. It can be understood as directly concatenating the vectors R_V and R_T into the classifier for classification.
MARN-CA: This model removes the cross-attention residual module. The R_V and R_T vectors are concatenated and then classified by self-attention residual fusion.
MARN-SA: This model removes the self-attention residual module. The vector R'_V and R_T are concatenated and classified.
MARN-Residual: This model removes the residual connection in the cross-attention residual and self-attention residual modules (the dotted line in **Figure 1**), and all other aspects remain unchanged.

The classification results of each ablation model are shown in **Table 4**.

From **Table 4**, it can be seen that when the cross-attention residual module and self-attention residual module are deleted at the same time, the accuracy of the model is at least 0.8359, but it is still higher than that of the textual model and all baseline models. There are two reasons for this result. One is that we used the Bert model with better text extraction, and the other is that the overall results are improved by the image features.

When the model removes the cross-attention residual module, the accuracy of the model is slightly lower than that of the overall model, which shows that it is effective to use the text feature to strengthen the image by using the cross-attention residual mechanism. The enhanced image can give greater weight to the key areas related to the text so that the features can be selected to deal with. Similarly, when the self-attention residual module is removed, the model results are lower than that of the overall model. This is because the cross-attention residual module only enhances the image and does not involve the text mode. The self-attention residual module enhances the attention of the text feature and image feature at the same time and further improves the classification performance of the model.

It is worth noting that when we remove the residual connection between the two modules, the model performance is also reduced, which is easier to understand. The function of the residual mechanism is to prevent the overall model performance from being worse than the original model. However, it can be seen from **Table 4** that the role of the residual mechanism does not stop there. It can make full use of the complementarity of multimodal information while keeping the unique attributes of each mode as much as possible, avoiding the adverse effects of noise information.

CONCLUSION

In this article, we propose the MARN model to solve the problem of insufficient feature fusion between modes and serious information redundancy after fusion. The model uses the multilevel attention residual module to fuse text and image features selectively. On the basis of making full use of each mode feature, the noise information generated during mode fusion is minimized, to a certain extent, resulting in the above two problems being solved. The experimental results show that the performance of the MARN model is better than the related baseline models and ablation models in terms of accuracy and F1 value. To the best of our knowledge, there is no research on video rumors. We are going to collect short video rumors and explore them, so as to expand the application scope of rumor detection.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/wangyajun-ops/Weibo-dataset.

AUTHOR CONTRIBUTIONS

WZ: responsible for article conception, model building, code implementation, and writing; SJ: corresponding author, responsible for guidance and revision work.

FUNDING

National Key Research and Development Program of China, No.2017YFB0803001; National Natural Science Foundation of China, No.61572459.

REFERENCES

1. O'reilly T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *MPRA Paper* (2007) 97(7):253–9.
2. Jin Z, Cao J, Zhang Y, Zhou J, and Tian Q. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Trans Multimedia* (2017) 19(3):598–608. doi:10.1109/TMM.2016.2617078
3. Alport GW, and Postman L. *The Psychology of Rumor*. Henry Holt: public opinion quarterly (1947). p. 45p.
4. Li S, Zhao D, Wu X, Tian Z, Li A, and Wang Z. Functional Immunization of Networks Based on Message Passing. *Appl Maths Comput* (2020) 366: 124728–8. doi:10.1016/j.amc.2019.124728
5. Huang K, Li S, Dai P, Wang Z, and Yu Z. SDARE: A Stacked Denoising Autoencoder Method for Game Dynamics Network Structure Reconstruction. *Neural Networks* (2020) 126:143–52. doi:10.1016/j.neunet.2020.03.008
6. Li S, Jiang L, Wu X, Han W, Zhao D, and Wang Z. A Weighted Network Community Detection Algorithm Based on Deep Learning. *Appl Maths Comput* (2021) 401(2021):126012–9. doi:10.1016/j.amc.2021.126012

7. Huang K, Wang Z, and Jusup M. Incorporating Latent Constraints to Enhance Inference of Network Structure. *IEEE Trans Netw Sci Eng* (2020) 7(1):466–75. doi:10.1109/TNSE.2018.2870687
8. Zhang P, Ran H, Jia C, Li X, and Han X. A Lightweight Propagation Path Aggregating Network with Neural Topic Model for Rumor Detection. *Neurocomputing* (2021) 458(2021):468–77. doi:10.1016/J.NEUCOM.2021.06.062
9. Castillo C, Mendoza M, and Poblete B. Information Credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web; 2011 March 28 - April 1; Hyderabad, India (2011) p. 675–84. doi:10.1145/1963405.1963500
10. Yang F, Liu Y, Yu X, and Yang M. Automatic Detection of Rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics; 2012 August 12; Beijing, China. New York, NY: MDS (2008) doi:10.1145/2350190.2350203
11. Mohammad SM, Sobhani P, and Kiritchenko S. Stance and Sentiment in Tweets. *ACM Trans Internet Technol* (2017) 17(3):1–23. doi:10.1145/3003433
12. Zhao Z, Resnick P, and Mei Q. Enquiring Minds. In: Proceedings of the 24th International Conference on World Wide Web; 2015 May 18-May 22; Florence, Italy (2015) doi:10.1145/2736277.2741637
13. Zhang X, Chen F, and Huang R. A Combination of RNN and CNN for Attention-Based Relation Classification. *Proced Comp Sci* (2018) 131:911–7. doi:10.1016/j.procs.2018.04.221
14. Liu Y, Jin X, Shen H, Bao P, and Cheng X. A Survey on Rumor Identification over Social Media. *Chin J Comp* (2018) 41(07):1536–58. doi:10.11897/SP.J.1016.2018.01536
15. Ma J, Gao W, Mitra P, Kwon S, Jansen B, Wong K, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence; 2016 July 9 - July 15; San Francisco, USA (2016). p. 3818–24.
16. Liu Z, Wei Z, and Zhang R. Rumor Detection Based on Convolutional Neural Network. *J Comp Appl* (2017) 37(11):3053–6. doi:10.11772/j.issn.1001-9081.2017.11.3053
17. Chen T, Li X, Yin H, and Zhang J. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. In: Proceedings of the Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference; 2018 June 3- June 6; Melbourne, Australia (2018) p. 40–52. doi:10.1007/978-3-030-04503-6_4
18. Chen Y, Sui J, Hu L, and Gong W. Attention-Residual Network with CNN for Rumor Detection. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 2019 November 3 - November 7; Beijing, China (2019) doi:10.1145/3357384.3357950
19. Poria S, Cambria E, Howard N, Huang G-B, and Hussain A. Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content. *Neurocomputing* (2016) 174:50–9. doi:10.1016/j.neucom.2015.01.095
20. Huang F, Zhang X, Zhao Z, Xu J, and Li Z. Image-text Sentiment Analysis via Deep Multimodal Attentive Fusion. *Knowledge-Based Syst* (2019) 167:26–37. doi:10.1016/j.knosys.2019.01.019
21. Jin Z, Cao J, Guo H, Zhang Y, and Luo J. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In: Proceedings of the 25th ACM international conference on Multimedia; 2017 October 14 - October 19; California, USA (2017) p. 795–803. doi:10.1145/3123266.3123454
22. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, and Gao J. Eann. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 July 19 - July 23; London, England (2018) doi:10.1145/3219819.3219903
23. Dhruv K, JaiPal S, Manish G, and Varma V. MVAE: Multiimodal Variational Autoencode for Fake News Detection. In: Proceedings of the 19th World Wide Web Conference; 2019 May 10 - May 14; San Francisco, USA (2019). doi:10.1145/3308558.3313552
24. Liu J, Feng K, Jeff Z, Juan D, and Wang L. MSRD: Multi-Modal Web Rumor Detection Method. *J Comp Res Dev* (2020) 57(11):2328–36. doi:10.7544/issn1000-1239.2020.20200413
25. Devlin J, Chang M, Lee K, and Toutanova K. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding* (2018). p. 04805. arXiv:1810.
26. He K, Zhang X, Ren S, and Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition; 2016 June 27 - June30; USA. Piscataway: NV (2016) p. 770–8. doi:10.1109/CVPR.2016.90
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention Is All You Need. *Adv Neural Inf Process Syst* (2017) 12:5999–6009.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.