Multimedia Tools and Applications (2022) 81:13799–13822
https://doi.org/10.1007/s11042-022-12290-8

An effective strategy for multi-modal fake news detection

Xu Peng[1] · **Bao Xintong**[1]

Received: 6 October 2020 / Revised: 24 November 2021 / Accepted: 14 January 2022 /
Published online: 24 February 2022

Abstract
News plays an indispensable role in the development of human society. With the emergence of new media, fake news including multi-modal content such as text and images has greater social harm. Therefore how to identify multi-modal fake news has been a challenge. The traditional methods of multi-modal fake news detection are to simply fuse the different modality information, such as concatenation and element-wise product, without considering the different impacts of the different modalities, which leads to the low accuracy of fake news detection. To address this issue, we design a new multi-modal attention adversarial fusion method built on the pre-training language model BERT, which consists of two important components: the attention mechanism and the adversarial mechanism. The attention mechanism is used to capture the differences in different modalities. The adversarial mechanism is to capture the correlation between different modalities. Experiments on a fake news Chinese public dataset indicate that our proposed new method achieves 5% higher in terms of F1.

Keywords Fake news · Multi-modal · BERT · Attention mechanism

1 Introduction

Recently, the Internet has gradually become the main way for people to obtain information. As the amount of information disseminated on the internet has increased, so has the emergence of fake news. According to previous work of [26], fake news is defined as deliberately fabricated and verified as false. Fake news jeopardises the security of the global Internet, and its scale, speed of dissemination and fraudulent methods are all growing rapidly. Gartner, an international consulting company, predicted that by 2020, fake news on the Internet

✉ Xu Peng
xupeng@bupt.edu.cn

Bao Xintong
bxt@hljdx.net

[1] State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Room 519, New Science Research Building, No. 10, Xi Tu Cheng Road, Beijing, China

will be rampant [5]. Therefore, to counter the proliferation of fake news and minimise its damage, there is a great need for in-depth research into automatic monitoring methods for fake news.

In the fake news detection task, the main problem is how to distinguish fake news according to its characteristics including sources, texts, attached pictures, etc. Based on these characteristics, there are two kinds of methods for fake news detection: single-modal and multi-modal, both of which are a classification task employing different features. In the detection of single-modal fake news, a single feature such as text or image is commonly used alone for news classification. In the detection of multi-modal fake news, combinations of different features need to be considered, while the complementarity of different modalities can eliminate the redundancy between them. Compared with the single-modal fake news detection method, the method of feature combination used in multi-modal fake news detection can provide a better feature representation of fake news to improve the accuracy. However, the traditional multi-modal fusion strategy also suffers from certain shortcomings in that it only uses simple fusion strategies [12], such as concatenation and element product, which do not yield a better feature representation. Therefore, it is essential to design a feature fusion strategy to detect the fake news.

In details, there are three challenges for multi-modal fake news detection task:

– *How to design a strategy to capture the different contributions of different modalities?*
  Different modalities pay different contributions to representation of all features. It is very important to design a special strategy to get better representation, because representation must include the feature of every modality.
– *How to design a strategy to capture the correlation of different modalities?*
  In addition to different contributions, different modalities also have correlations. The correlation of different modalities represents the common feature which is the part of the representation.
– *How to design a more effective multi-modal fusion network to detect fake news?*
  The disadvantage of traditional strategy is that they just project different modality space to the same modality space, but do not consider fuse the different contributions and correlation of them.

In order to address the aforementioned challenge, we study how to better integrate different modality features of fake news detection. First of all, we analyze some fake news data and conclude that in the fake news detection, although text feature is the main feature, other modality features are also important and make contributions. Then, we design a new network called Multi-modal Attention Adversarial (MMAA) fusion network to integrate text and other modality features and some experiments to explore the details of the strategy. Verified by experiment, the MMAA achieved good results for fake news detection on a fake news Chinese public dataset. In the MMAA, we use the attention mechanism to capture the differences between different modalities and use the adversarial mechanism to capture the correlation of different modalities.

The contribution of this paper is as follows:

(1) As far as I know, this is the first time to attempt a multi-modal fake news detection method that combines the attention mechanism and the adversarial mechanism.
(2) We design a Multi-modal attention mechanism to capture the differences of different modality feature, then we design a Multi-modal ad versarial mechanism to capture

the correlation of different modality features and finally we design the Multi-modal Attention Adversarial network to detect the fake news.

(3) We have conducted extensive experiments on a fake news Chinese public dataset. They show that our proposed approach achieves 5% better than the traditional method in terms of F1.

2 Related works

According to the summary of [33], there are two kinds of methods for fake news detection: single-modal-based and multi-modal-based.

Fake news detection based on single modality: There are several features for fake news including text, image, social content, and category. The single-model-based method only makes use of one feature to do the detection. Each feature is described as follows.

The text feature is extracted from the text of the news. However, the features extracted by traditional methods cannot represent the internal meaning of language well, because they are highly dependent on facts and domain knowledge. Therefore, it is difficult to use the traditional machine learning model to detect fake news. To overcome this problem, [17] proposes a deep learning model to learn language representation and mainly use Recurrent Neural Network (RNN) to extract text features. [23] exploits the linguistic features of misinformation by comparing real news with fake news. In literature [19], the authors predict the stance of a set of texts representing facts with respect to a given claim by using end-to-end memory networks.

The image feature is also important. For fake news detection. There are some methods bases on image features such as [12]. However, the features extracted in these methods cannot represent the complex distribution of images and the content of fake news well.

The social context feature expresses the characteristics of users who publish information on social media such as the gender of the user, the number of friends, and the number of followers. In general, social features are not used alone for fake news detection, but employed as additional features to provide supplementary information [8]. This is due to the fact that there is not enough information in the social features to detect fake news.

The category feature expresses the category of news. At present, there is no specific method based on category features to detect fake news. Like social context features, it is also used as additional features to help fake news detection.

Overall, the text feature and the image feature can be used to detect fake news independently, while other features are usually used as the additional features to help the detection.

Fake news detection based on multi modality: Multi-modal fake news detection belongs to multi-modal tasks. In multi-modal tasks, existing work includes but is not limited to [2], image Capturing [3, 14] conducts fake news detection by evaluating theconsistency between the body and its claim given a news article, this paper [35] proposes a RNN with an attention mechanism to fuse multi-modal data from tweets for rumor detection. The key to the multi-modal task is to find a good multi-modal representation and deep neural network has shown good results in learning different representations in these tasks, in which the pre-training model has played a great role. The text modality and the image modality are also used

in the multi-modal tasks. This study [36] proposes a machine learning-based Similarity-Aware FakE news detection method (SAFE) for extracting and analyzing the relationship between textual and visual information in news articles. The study [29] proposes SpotFake+ which is a multi-modal approach. And SpotFake+ uses transfer learning to obtain semantic and contextual information from the text and images of news for fake news detection and achieves good accuracy. This study [25] proposes a cultural algorithm using situational and normative knowledge to detect fake news containing text and images. This study [30] presents SpotFake-, a multi-modal (textual and visual features) framework for fake news detection. It enables the detection of fake news without regard to any other subtasks. The study [16] presents a Multimodal Variational Autoencoder (MVAE) that combines a bimodal variational autoencoder and a binary classifier to perform the fake news detection task. The model mainly obtains detection results by combining textual and visual information. The study [33] proposes an end-to-end framework of Event Adversarial Neural Network (EANN), which enables fake news detection by extracting textual and visual features, removing event-specific features, and retaining shared features between events. In addition to the text modaloty and the image modality, there are existing works that propose the use of information from other modalities for fake news detection. The study [34] proposes a combined textual information, knowledge concepts and visual information, knowledge-driven Multimodal Graph Convolutional Network (KMGCN) for fake news detection. This study [31] proposes a multi-modal fake news detection framework which is based on Crossmodal Attention Residual and Multichannel convolutional neural Networks (CARM-N). And CARM-N is mainly implemented to maintain the unique information of each modality while fusing the relevant information between different modalities, as well as to mitigate the impact of the noise information that may arise from the crossmodal fusion component. For different modalities, many different pre-training models are helpful to get single-modal representation.

In the modal of computer vision, there are many popular pre-training models, such as VGG, ResNet, etc. They generally refer to the CNN trained on the ImageNet dataset that contains more than 14 million images and 1000 categories for helping to train l models.

In the modal of natural language processing, many models are used to get word vector, such as Word2vec, glove [21], cove [18], etc. Although they are very useful and trained in a large corpus, they cannot capture the relationship between words by using word vectors as pre-training models. With the development of deep learning, the natural language pre-training model solves this problem well and Bert [7] is the most popular natural language pre-training model.

After using different pre-training models to get text and image representations, we need to fuse them. These papers [12], use deep learning to build fake news multi-modal model. However, in the process of modeling, these papers only use the simple fusion strategy to fuse multi-modal features and do not consider that different categories of news pay different attention to different modals. In order to overcome the shortcomings of existing work, we propose a deep learning model, which uses the attention mechanism to dynamically adjust the model's attention to get different modal information.

3 Methodology

Fake news detection is a multi-modal classification task, a survey on different content types of news and their impacts on readers can be found in [20]. Text feature is the main feature

and the others are used as the additional information, so we choose the Bert as the core model to get text representation and fuse it with other modality features.

In this paper, we study the strategy of integrating different modality features of fake news detection. In order to address this problem, the proposed model should satisfy three requirements: (1) The main feature is text feature for fake news task, the better text representation is very important. (2) Capture the differences of different modality features; (3) Capture the correlation of different modality features.

To solve the problem, there are some factors: (1) Bert for extracting text representation. (2) Attention mechanism for capturing the differences of different modality features. (3) Adversarial mechanism for capturing the correlation of different modality feature. (4) The Multi-modal Attention Adversarial (MMAA) network can be used to solve multi modality task.

Extract Text Representation The paper [32] uses the Bert for single modal text classification finetune and it is not difficult to get text representation. The Bert model contains twelve transformer layers with a hidden size of 768, and the inputs of it are 512 tokens and the output is the representation of the whole sentence. The input sentences will be classified into two special tokens, CLS and SEP, which are used to do the classification and separate different paragraphs respectively. In general, a simple softmax layer is added to the output of Bert to predict labels. We can use the same method to get text representation and fuse it with other features sequentially. This chapter will introduce how to use the MMAA strategy for multi-modal fake news detection, which is a better fusion strategy for the Bert in multi-modal text classification.

3.1 Attention mechanism

The attention mechanism is actually an addressing process. As shown in the Fig. 1, given a task-related query vector Q, the attention value is calculated by calculating the attention distribution with key and attaching it to the value. This process is actually the embodiment of the attention mechanism to reduce the computational complexity: it is not necessary to input all information into the neural network for calculation, but only select some task-related information.

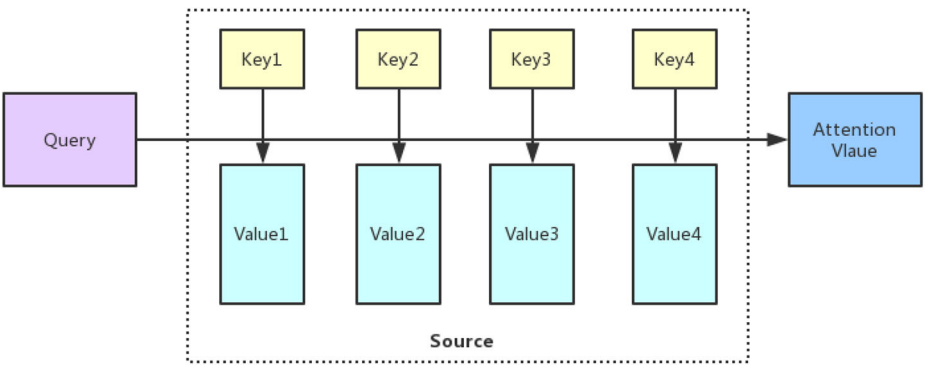

Fig. 1 Soft attention

The specific mathematical form is as follows:

$$Attention(Query, Source) = A \tag{1}$$

$$A = \sum_{i}^{L_x} similarity(Query, Key_i) Value_i \tag{2}$$

The calculation process is divided into three stages:

(1) In the first stage, calculating the similarity or correlation between query and key. The most common methods include:

- **Addition** $s(x_i, q) = v^T tanh(Wx_i + Uq)$
- **Product** $s(x_i, q) = x_i^T q$
- **Bilinear** $s(x_i, q) = x_i^T Wq$

(2) In the second stage, with the normalization operation, the weight of important elements can be more prominent through the internal mechanism of softmax. That is, the following formula is generally used for calculation:

$$a_i = \frac{e^{sim_i}}{\sum_i^{Lx} e^{sim_i}} \tag{3}$$

(3) In the third stage, weighted summation can obtail the value of attention.

$$Attenton(Query, Source) = \sum_{i}^{L_x} a_i V_i \tag{4}$$

Figure 2 shows the architecture of fake news multi-modal attention mechanism. In fake news detection, different categories of fake news contribute differently modality features. Therefore, category feature can be used as the query vector while other features can be used as the key vectors. Because key vectors are the same as the value vectors, other features can be used as value vectors.

According to the attention mechanism, the query vector is used to calculate the similarity with other features and the similarity can be used to calculate with the value vector. Finally, all vectors are added together as the final representation to make results.

3.2 Adversarial mechanism

The attention mechanism explores the dynamic contribution of modality uniqueness while ignoring the correlation between multi modalities, which is an important component of multi modality representation. Therefore, we consider learning the correlation between multi modality by introducing an adversarial mechanism into the model we design, which in turn further improves the uniqueness of the modality uniqueness features. Since the contribution of the modality uniqueness features is computed dynamically, it is reasonable to give fixed weights to the modality invariant features by a multi-modal adversarial network.

An effective shared subspace is obtained by the interaction between different modality of the adversarial learning mechanism that is based on the interaction between two different processes.

(1) The first process is a feature mapper, which attempts to produce modalities invariant representations in the shared subspace and confuse another process.

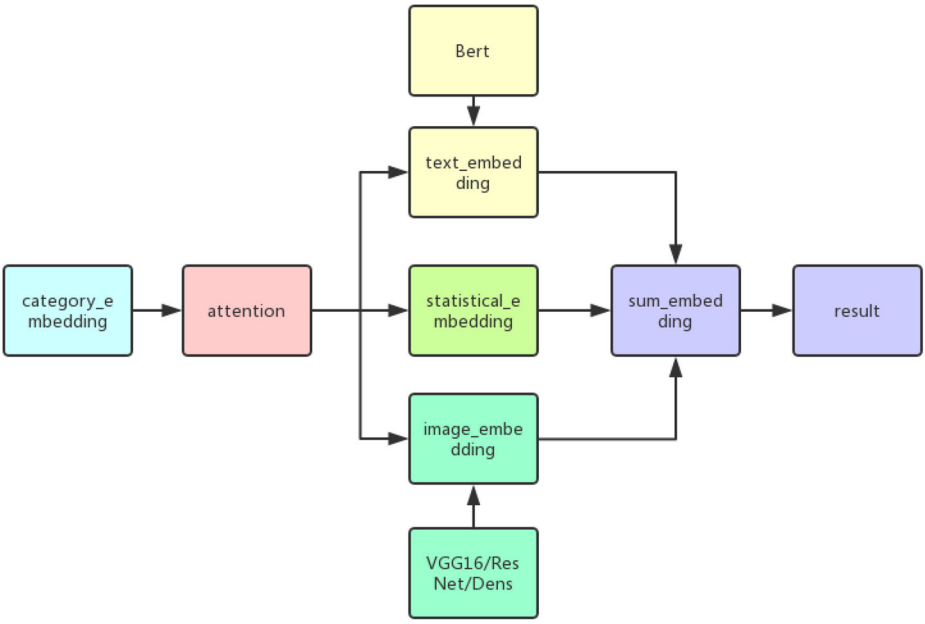

Fig. 2 Multi-Modal attention mechanism

(2) The second is a modality classifier, which distinguishes different modalities based on the feature representation information generated in the first process.

Adversarial mechanism is built around the mini-max game and the objective function is as follows:

$$minmaxL_D = E_{x\sim p_{x1}}[log(D(I(x)))] + E_{x\sim p_{x2}}[log(1 - D(I(x)))] \tag{5}$$

Among them, p_{x1} and p_{x2} distributions of two modalities, modality classifier D classify different modalities, and feature mapper I produce invariant representations to confuse modality classifier. For the fixed feature mapper, the optimal solution of modality classifier is

$$D^*{}_I(x) = \frac{p_{x1}(x)}{p_{x1}(x) + p_{x2}(x)} \tag{6}$$

The above-mentioned method focuses on the confrontation paradigm. However, fake news involves multiple different modalities and different modality features contain different degrees of shared potential features, so we can learn the degree of migration of each modality to guide the modality classifier.

To solve the problem, we introduce two classifiers. The first classifier C_0 outputs the degree of migration of each modality and the second classifier C_1 output the probability of modalities invariant representations belonging to the right modality.

The first classifier C_0 is a three-class classifier and the detail formula is as follows:

$$e = [e_1; e_2; e_3] \tag{7}$$

$$C_0^i(x) = softmax(I(e)) \tag{8}$$

where e_i represents the embedding of different modality, I represents feature mapper, $I(e_i)$ represents the invariant features, and C_0^i represents the probability of invariant feature i belonging to the right modality. If $C_0^i \approx 1$, $I(e_i)$ is impossible to include invariance features shared by other modes, because it can be completely distinguished from other modes by the discriminator. Therefore, the contribution of degree to the mode invariant feature should be inversely related, as follows:

$$w_i(x) = 1 - C_0^i(x) \tag{9}$$

The second classifier C_1 is also a three-class classifier. By adding degree, the detail formula of objective function is as follows:

$$minmaxL_D = E_{e\sim p_{e^i}} \sum_{i=1}^{3} [w_i(x) log(C_1(I(e)))] \tag{10}$$

Among them, $w_i(x) = w_i(I(e))$ is a constant, independent of D_1, for the fixed I and D_0, the optimal output is:

$$D_1^{i*}(x) = \frac{w_i(x) px_i(x)}{\sum_{i=1}^{3} w_i(x) px_i(x)} \tag{11}$$

when $w_1(x)px_1(x) = ... = w_3(x)px_3(x)$, formula(2) can get the optimal value. As a result of $px^m(x) = c^m = I(e^m)$, we can get the representation of the invariance of the weighted modality invariance by max-pooling, which means taking the maximum value. $c_{max} = max(w_i c_i)$ is used to represent invariant feature. The Fig. 3 shows the architecture of fake news multi modality adversarial mechanism.

3.3 MMAA network

According to the above methods, the features of multi-modal fake news detection include text features, image features, statistic features, category features, etc. In this case, the main task is to better fuse different features together.

In the fake news detection, different modes are emphasized for different news categories. For some news categories, the text features are more important than the image features that only play an auxiliary role, while for some categories the image features play a more important role. Therefore, model need to consider characteristics of different modalities dynamically according to different news categories. In addition, attention mechanism captures the dynamic contribution of different and ignores the correlation of multi-modal which is also the important part of multi-modal. Based on such considerations, the core of MMAA is designed to use attention mechanism to dynamically adjust weight of different modality and uses adversarial mechanism to capture the invariant modality feature. The Fig. 4 shows the architecture of MMAA strategy.

MMAA can help to fuse different modality features, but there are different considerations for each single-modal feature.

For Text Feature: Bert model is composed of 12 transformer layers. After the initial input of 512 tokens, segment embedding and position embedding will be added together as the input of transformer. When Bert processes text, some extra text features are ignored, such

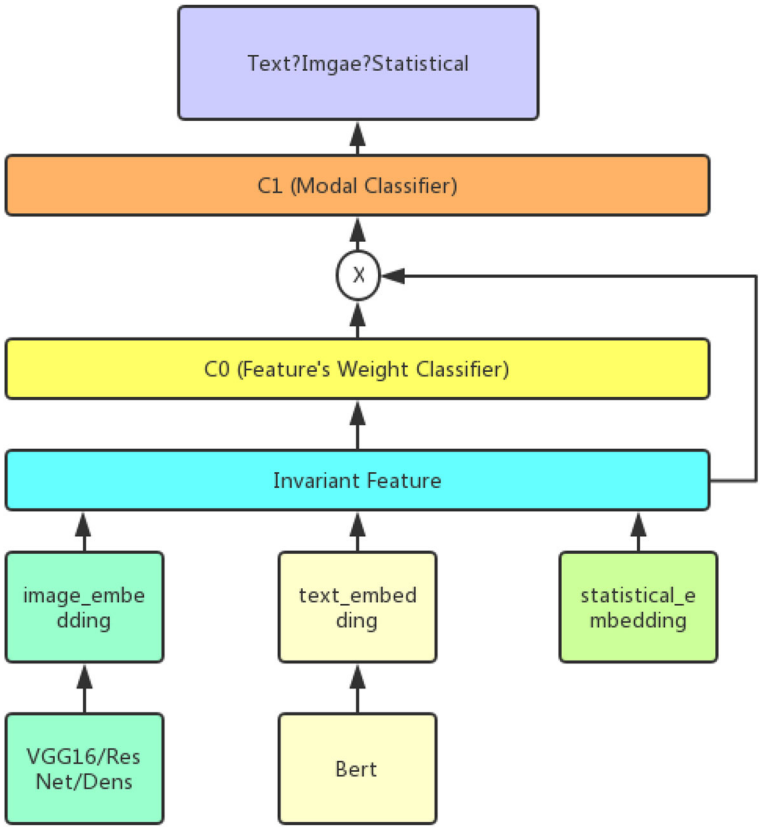

Fig. 3 Multi-Modal adversarial mechanism

as the size and the color of the text, no matter the text is a symbol or a text. Those additional text features will further improve the performance and robustness of the model. There are two methods to fusion extra text features:

1) Adding the extra embedding, segment embedding and positional embedding together as the whole input before being inputted into transformer.
2) Adding extra embedding after the output of Bert, and using concatenate or element-wise sum.

For the fake news detection task, there are some extra text features can be used and the details can be introduced in the experiment section.

For Image Feature: Traditional multi-modal tasks, such as VQA tasks, can be divided into two types to fuse image features.

The first type is using the image pre-training model to extract the image features and fusing them with text features together. There are some methods about this type: [1, 13], which combine the image and text features using simple mechanisms such as concatenation, elementwise multiplication, or elementwise addition, and then processing them with a linear classifier or a neural network.

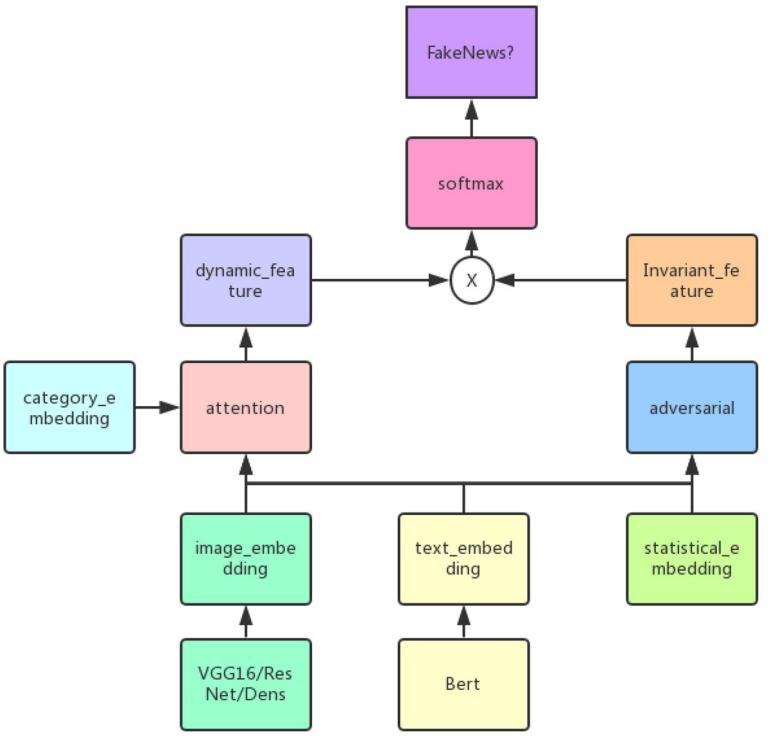

Fig. 4 Multi-Modal Attention Adversarial

The second type is using FRCNN [24] to extract the region features of the image, and then using the attention mechanism to integrate the image and text features, in which the classifier uses the text features to compute spatial attention maps for the image features or adaptively scales local features based on their relative importance. This kind of method does not work well for the fake news detection task, because image region feature of fake news data plays an insignificant role. So, in this paper, the image pre-training model is used to extract the whole image features and then the fusion strategy is used to fuse them.

For Statistic Feature: Statistic features are always used as the additional information because statistic features are sparse in the data. The data preprocessing will first one-hot the statistical features, and then input them as a whole vector to fuse with the text features. For the fake news detection, the gender of the user, the number of friends, and the number of followers are the statistical features can be used.

4 Experiments

Our proposed fusion strategy MMAA fuses three modality features including text feature, image feature, and statistical feature. We choose a multi-modal Chinese fake news dataset containing varying numbers of documents and different lengths of documents that includes four modalities of data: text, image, statistic, and category. The fusion strategy MMAA was evaluated on this dataset and compared with traditional fusion strategies.

We chose the Chinese BERT-base classification model as the baseline model, in which the original data after simple pre-processed is first fed into the Chinese BERT-base classification model, and then the classification results are output. Based on this idea, we conduct four experiments based on the baseline model.

The first experiment shows the evaluation results of the MMAA strategy on the dataset and some ablation studies are done to indicate the effectiveness of our strategy better. The other experiments show the result of the strategy by merely using the main text feature and other features. The second experiment shows the result of the case that only the text feature and the extra text feature are used and the third experiment is about the case that only the text feature and the image feature are applied. The last experiment shows the result of the fusion strategy only using the text feature and the statistical feature. Meanwhile, these last three experiments show how to get a better representation of different modalities.

Here is a brief introduction to each experiment:

(1) **Exp-I: MMAA Strategy For All features**: Evaluating the MMAA strategy on multi-modal fake news dataset.
(2) **Exp-II: Strategy For E xtra Text Feature**: Exploring the effect of adding the extra text feature.
(3) **Exp-III: Strategy For Image Feature**: Exploring the effect of fusing the image feature.
(4) **Exp-IV: Strategy For Statistical Feature**: Exploring the effect of fusing the statistical feature.

4.1 Datasets

All the experiments are carried out in the multi-modal Chinese fake news dataset[1] containing varying numbers of documents and different document lengths.

Multi-Modal Chinese Fake News In the multi-modal Chinese fake news dataset, there are four modalities of data: text, image, statistic, and category. The specific data fields are shown in Table 1. There are 16348 texts and 16348 corresponding images in the train datasets, including 11064 real news with 11064 corresponding images and 5284 fake news with 5284 corresponding images. There are 1551 texts and 1551 images in the test dataset, which has the same proportion of real and fake news as that of the train set.

4.2 Data preprocessing

In the fake news dataset, there are text, image, statistics, and seven categories of news. For different modality features, there are different preprocessing methods.

4.2.1 Process extra text data

There are several steps in the preprocessing of the extra text data:

– Designing customed stop words list and use word segmentation tools which named jieba to segment text.
– Using the TF-IDF algorithm trained on extra data to get the TF-IDF vector.

[1] https://github.com/1210882202/data

Table 1 Data Fields

Name of fields	Fields	Detail
id	num	news id
content	text	content of news
piclist	url	url of image
gender	num	gendr of user
follow count	num	number of user followers
fans count	num	number of user fans
location	text	location of user
description	text	description of user
weibo count	num	number of user weibo
category	text	category of news
label	num	label of news

- Sorting the TF-IDF vector and getting the key-words of the extra data.
- Judging each word in the train data. If the word in the train data is a keyword, it is marked as 1. If not, it is marked as 0.

TF-IDF: TF-IDF is a statistical method to evaluate the importance of a word to a document set or one of the documents in a corpus. The importance of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the number of times it appears in the corpus. The following is an introduction to the specific formula:

$$TF - IDF = TF \times IDF \tag{12}$$

TF (term frequency) means word frequency, which means the number of times a word appears in an article. The following is the definition of TF_w:

$$TF_w = \frac{C_w}{C} \tag{13}$$

where C means all words in an article and w means the number of keywords in an article.

IDF (inverse document frequency) is the inverse text frequency index. If fewer documents contain the keyword w, the key-word w has a better ability to distinguish categories. The following is the calculation of IDF_w:

$$IDF_w = log\frac{C}{C_w + 1} \tag{14}$$

4.2.2 Proess image data

In image preprocessing, each image is scaled to 224*224 and each pixel is normalized. As there are some training data without images, a 224*224*3 zero matrix is created manually and used as the image for these data. The VGG16 [27], ResNet50 [10], Densenet121 [11] are chosen as the image pre-training models to extract image features. There are brief introductions about three models.

VGG16 [27]**:** VGG is a convolutional neural network model proposed by Simonyan and Zisserman in the literature" very deep convolutional networks for large scale image recognition", and it shows very good results in image classification and target detection tasks [28].

There are some variants of the VGG model, but the most popular one is VGG16 consisting of 16 layers.

ResNet50 [10]**:** ResNet is the best paper published on CVPR2016. The most fundamental motivation of ResNet is the so-called degenerate problem that when the layer of the model is deepened, the error rate is increased. To solve this problem, ResNet proposes a residual block structure, and each residual block uses a short cut connection. The simple addition does not add extra parameters and computation to the network, but greatly increase the training speed and improve the training effect of the model. When the number of layers of the model is deepened, the simple structure can solve the degradation problem well (Fig. 5).

Densenet121 [11]**:** Densenet (Dense connected progressive networks) is the best paper of CVPR2017. Based on the core idea called skip-layer, the author of Densenet designs a new connection mode. To maximize the information flow between all layers in the network, the model connects all layers, so that each layer accepts the characteristics of all layers in front of it as input. Densenet mainly has the following two characteristics (Fig. 6):

– To some extent, it can alleviate the problem of gradient dissipation in the training process. In the structure, each layer receives the gradient from all subsequent layers during backpropagation, so the gradient near the input layer does not become smaller with the increase of network depth.
– Because a large number of features are reused, a large number of features can be generated by using a small number of convolution kernels, and the size of the final model is relatively small.

4.2.3 Process statistic data

In the multi-modal fake news datasets, the user-related data are all statistical data. To apply the statistical data to the training, the continuous data is firstly discretized. For different statistical data, the way of discretization is also different. The statistical data in the fake

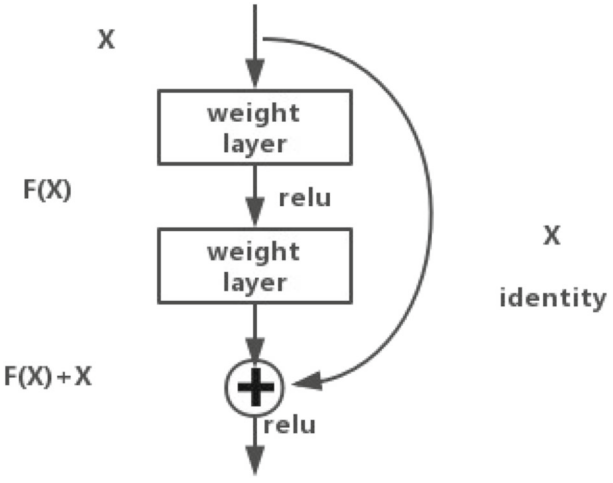

Fig. 5 Residual block

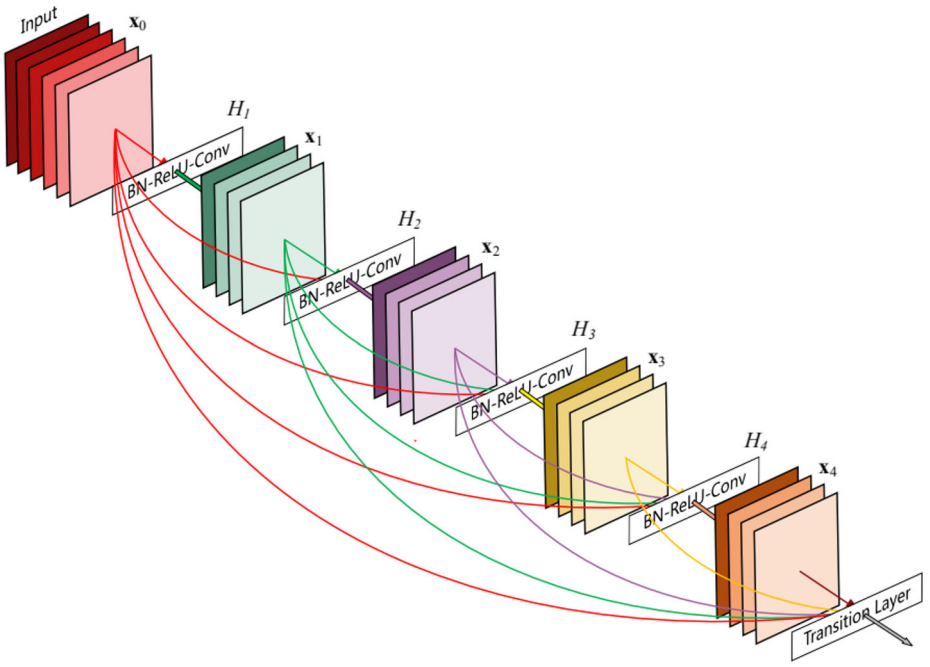

Fig. 6 Densenet block

news data include user's gender, user's Weibo fans, user's followers, and the number of user's Weibo.

- **User's gender** Male is considered as 0, female as 1 and nan as -1.
- **User's weibo information** Dividing the corresponding intervals according to the maximum and minimum values of numbers, and discretize different intervals into fixed numbers.

We use the Chinese BERT-base model with a hidden size of 768, 12 Transformer layers, and 12 self-attention heads. The BERT is further fine-tuned on 2 TITAN Xp GPU with the batch size of 32, the max sequence length of 256, the learning rate of 3e-5, the train epochs of 3 and the warm-up steps of 1000. The dropout probability is always kept at 0.1. We use Adam with β_1=0.9 and $\beta_2 = 0.999$. The max number of the epoch is empirically set to 3 to save the best model on the validation set for testing.

4.3 Compared methods

As the same as the [6], we compare SAME with several representative and stateof-the-art fake news detection methods including KNN, SVM, EANN [33] and CSI [22].

- KNN: This determines the authenticity of news based on the labels of its neighbors.
- SVM: We concatenate the features including the outputs of VGGNet, GloVe and one-hot encoding, and sentiment polarity distribution vector as the input of Linear.
- EANN: In this method, both text and image information are taken into consideration. This method uses an event discriminator in order to eliminate the effects of the event-specific features and maintain the common features among all these studied

events. We remove the event discriminator of this method as our datasets do not have event labels.
– CSI: This method explores all of news content, users responses to the news, and the sources that users promote in detecting fake news. However, as our datasets do not have time interval information in users' comment, we modify the codes accordingly.

4.4 Exp-I: MMAA strategy for all features

The MMAA means multi-modal dynamic fusion, the core of which is using category features to dynamic adjust other weights of modality features. Category features refer to the different categories of data. The MMAA uses category features to calculate the attention between other modality features according to the calculation result, and decide the contribution of different modality features. This experiment shows the result of MMAA on the multi-modal fake news dataset.

Figure 4 shows the architecture of our model, which involves four modality features including category features, text features, image features, and statistical features. The category features are used for calculating attention scores with other features to dynamically adjust the contribution of other modality features. Besides, there are many options for other modality features. In the data preprocessing, we describe the extraction methods of different modal features in detail and get the most appropriate way to deal with each modality feature in Exp-II, Exp-III, and Exp-IV.

Therefore, according to the experimental results of the following experiments, we adopt the method of after output and element-wise addition for additional text features. The Densenet121 is applied to extract image features. According to the statistical characteristics, the data processing only proposed one method, which can be adopted directly. After determining the processing methods of different modality features, the experiment is evaluated on the dataset, and the experimental results are analyzed in the next section.

4.4.1 Experiment result

In this experiment, Table 2 shows the result of our strategy and Table 3 shows the result of all strategy. During the experiment, the baseline model represents the Chinese BERT-base classification model. The result of MMAA achieves about 5% higher f1 than the baseline. To better explore how the attention mechanism contributes, the attention map is built. Figure 7 shows the attention score for different modality features.

From the Fig. 7, the X-axis represents the category characteristics of news, the Y-axis represents the contribution percentage of modality features, and histogram with different colors represents different modality features.

Table 2 The result of MMAA strategy

Model	Precision	Recall	F1	Accuracy
KNN	0.7398	0.7118	0.7202	0.7556
SVM	0.7954	0.7638	0.7743	0.8021
EANN	0.8059	0.7957	0.8002	0.8188
CSI	0.8293	0.7714	0.7872	0.8188
Baseline	0.7897	0.7591	0.7693	0.7975
Baseline-MMAA	0.8555	0.8179	0.8310	0.8517

Table 3 The result of all strategy

Model	Precision	Recall	F1	Accuracy
Baseline	0.7897	0.7591	0.7693	0.7975
Baseline-Extra-Text	0.7876	0.7656	0.7737	0.7988
Baseline-Image	0.8374	0.8032	0.8152	0.8188
Baseline-Statistical	0.7821	0.7650	0.7717	0.8188
Baseline-MMAA	0.8555	0.8179	0.8310	0.8517

It can be intuitively seen that for different categories of news data, the weight of different modality is indeed different. The conclusion indicated by this figure can be analyzed from the perspective of human experience. For example, in the fields of business, statistical feature makes more contribution, since someone with more fans may public real news generally.

The Table 3 shows all the experiment results. Baseline-Extra-Text, Baseline-Image and Baseline-Statistical model correspond to Exp-II, Exp-III, and Exp-IV. Baseline-MMAA is our model, which achieves the best result.

The Tables 4 and 5 show a number of edge cases that our model correctly classified while being incorrectly classified by Baseline. Since the original dataset is in Chinese, the contents of the dataset are translated and presented in Table 5. In Table 4, the first column "Id" represents the id of the news item in the dataset, the second column represents the content of the news, and the last column represents the url of the image included in the news.

4.5 Exp-II: Strategy For extra text feature

Text features refer to the specific meaning of each word and different datasets have different extra text features. The extra data in the multi-modal Chinese fake news dataset include

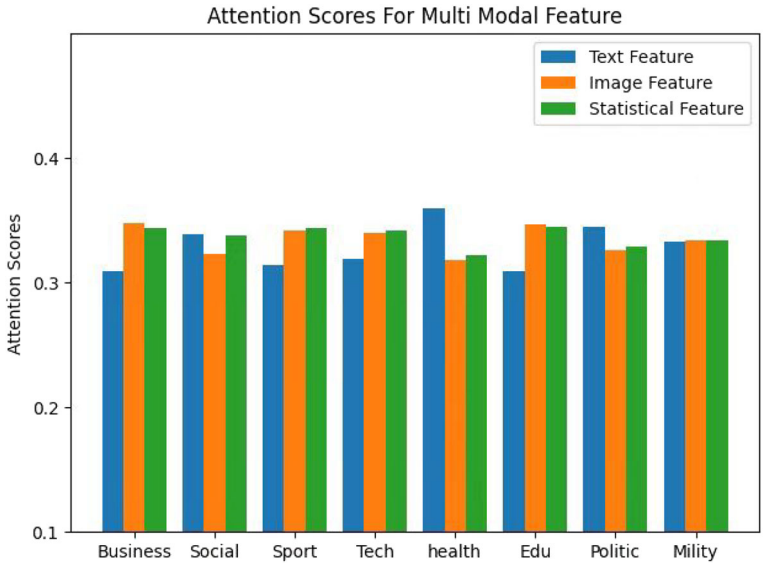

Fig. 7 Attentoin scores for multi-modal feature

Table 4 Edge cases

Id	Content	Piclist
f798623d3a9d91b6 8b3423b3191fbae6	在德国，汽车、公司都无需年检。 我问德国人：汽车不年检，坏了怎么办? 答：自己修车。 问：企业不年检，公司倒闭如潮怎么办? 德国人答：你看德国是这样的吗? 我问：德国政府为何不强制年检? 德国人反问：谁给了政府这个权力? 如果政府对年检感兴趣，说明这种事对它有好处。	b31bdca24918865db 56afa187391f657.jpg
9b292bf15f5c23b3 e5c5519dd3d09e6c	南无阿弥陀佛 @凤凰网华人佛教 @传喜法师.	9cdf66705731c7d04 d097f478183ce27.jpg

some real and fake news. Some keywords that are always appeared in the fake news are extracted by TF-IDF from the extra data and added into the model as text features, so that the ability of the model to recognize these words is strengthened. There are two methods of text feature fusion. One is to fuse text feature before transformer layers in Bert models and the other is to fuse text feature after Bert's output.

4.5.1 Fusion strategy

This experiment aims to explore how to fuse text features with extra embedding. After data pre-processing, every word of text has a symbol that is used to train extra text embedding and fused with text features.

There are two strategies to add extra text embedding. One is adding text embedding before the input of the transformer layer in Bert model, and the other is adding text embedding after the output of Bert model.

Table 5 Tranlation of edge cases

Id	Translation of content	Piclist
f798623d3a9d91b6 8b3423b3191fbae6	In Germany, cars and companies are not subject to annual inspection. I asked the Germans: What happens when a car breaks down without an annual inspection? Answer: Fix the car yourself. Q: What if a company doesn't have an annual inspection and it closes down like a tidal wave? German people answer: Do you see Germany is like this? I asked: Why does the German government not force annual inspection? The German asked in return: Who gave the government this power? If the government is interested in annual inspection, it means such thing is good for it.	b31bdca24918865db 56afa187391f657.jpg
9b292bf15f5c23b3 e5c5519dd3d09e6c	Namo Amitabha Buddha @Phoenix Chinese Buddhism @Venerable Chuan Hei.	9cdf66705731c7d04 d097f478183ce27.jpg

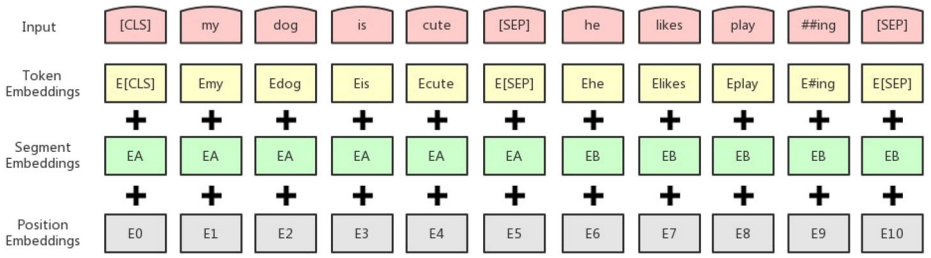

Fig. 8 The original input of Bert

Before Bert's input: The original input of the Bert model is shown in Fig. 8, including the sum of three embeddings including:

- **Token embedding**: indicate the embedding of the current word.
- **Segment embedding**: indicate the segment embedding of the sentence where the current word is belonged to.
- **Position embedding**: indicate the index embedding where the current word is located.

This strategy uses the extracted text features as the input of the Bert model in the form of embedding and the extra text embedding is added after position embedding. The new input of the Bert model is shown in Fig. 9. Text extra embedding indicates the keyword embedding of the sentence where the current word is marked, where 0 means that the current word is a keyword and 1 means not.

After Bert's Output: For different tasks, there are two text representation from the output of Bert. One is the first token embedding and the other is the whole embedding that includes every token embedding. If the task is a sequence level classification task, the output representation of the first token embedding is used to get the classification result output through a softmax layer, while if the task is the token level classification(such as NER), the output of the last layer of all tokens is chosen and then classified by the softmax layer. In this experiment, the first text representation is taken. So, the extra text embedding is fused with the first token output of Bert and then processed by a layer of transformer structure. Finally,

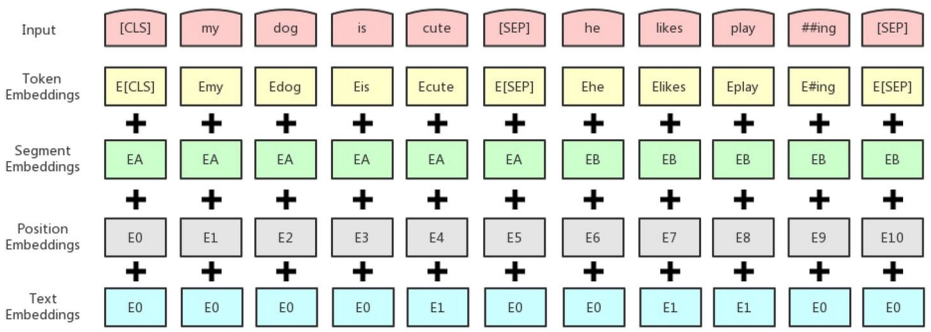

Fig. 9 The new input of Bert

the first token representation is used to get the final result from the softmax layer. The detail operation is shown as below:

$$first\_token = bert.get\_first\_token\_output() \tag{15}$$

$$extra\_text\_embedding = embedding\_look\_up() \tag{16}$$

$$ouput1 = first\_token + extra\_embedding \tag{17}$$

$$ouput2 = concat[first\_token, extra\_embedding] \tag{18}$$

There are two methods to fuse the first token embedding and the extra text embedding. The first method is element-wise addition and the second is concatenation.

Element-wise addition Expressed mathematically as: Existing feature vector $v_1 \in \mathbb{R}^n$, $v_2 \in \mathbb{R}^n$, to fuse the two feature vector, they are added directly in element-wise.

$$v = v_1 + v_2 \tag{19}$$

$$v = \{x_i \mid x_i = v_1[i] + v_2[i], i = 1, ..., n\} \tag{20}$$

The premise of this operation is that the dimensions of these two vectors are the same. If they have different dimensions, linear transformation can be used to transform them. $W \in n \times m$, $v \in \mathbb{R}^m$

$$v_n = Wv \in \mathbb{R}^n \tag{21}$$

Concatenation Concatenation is a more general feature fusion method. Its mathematical expression is as follows: Existing feature vector $v_1 \in \mathbb{R}^n$, $v_2 \in \mathbb{R}^n$, they are concatenated at the same dimension, and the result is fusion feature vector $v = [v_1, v_2] \in \mathbb{R}^{n+m}$

4.5.2 Experiment result

In this experiment, three sub-experiments are designed. The first sub-experiment is about the first fusion strategy that fuses the extra embedding before the input of Berts. The second and third sub-experiment are about the second fusion strategy that fuses extra embedding after the output of Bert. Table 6 shows the result of first sub experiment and Table 7 shows the result of others experimetns.

- **Before the input of Bert**: Fusing extra embedding before Bert's input.
- **After the output of Bert**: The strategy of fusing extra embedding is concatenated with the output of Bert.
- **After output of Bert**: The strategy of fusing extra embedding is added with the output of Bert.

From Table 6, the metric of the strategy fusing extra embedding before the input of Bert achieves improved recall and F1 scores with about the same accuracy as the original.

From Table 7, the metric of the strategy that uses concatenation to fuse the extra embedding with the output of Bert is achieves 0.6% higher F1 and 0.2% higher Accuracy.

Table 6 The result of first strategy

Model	Precision	Recall	F1	Accuracy
Baseline	0.7897	0.7591	0.7693	0.7975
Before Bert's input	0.7786	0.7634	0.7695	0.7930

Table 7 The result of second strategy

Model	Concat	Add	Precision	Recall	F1	Accuracy
Baseline	No	No	0.7897	0.7591	0.7693	0.7975
After output	Yes	No	0.7876	0.7656	0.7737	0.7988
After output	No	Yes	0.7864	0.7600	0.7692	0.7963

The conclusion of paper [15]: Addition is a special form of concatenation, which is a feature fusion method with prior knowledge added. In our experiment, element-wise addition and concatenation strategy have the similar result. The result of the concatenation is a little better than the element-wise addition, but the params of the two models the same bigger. In this case, the element-wise addition is chosen as the text fusion strategy.

4.6 Exp-III: Strategy For Image Feature

This experiment focuses on how to fuse the image feature with the text feature and test some fusion strategy.

The image feature can be extracted from the image pre-training model. There are many kinds of pre-training models and different models have different effects. Specifically, some representative image pre-training models are selected in this experiment, such as VGG16, ResNet50, and Densenet121. These image pre-training models will be fused with the text features by different fusion strategies.

The strategy of image feature fusion is different from that of text feature fusion. Text feature fusion uses concatenation or element-wise addition while image feature fusion uses concatenation, element-wise product, or MurelBlock. Murel-Block is a variation strategy of MCB [9]. The author of MCB thinks that these simple strategies are not as effective as the outer product and not enough to model the complex relationship between two modalities. But the complexity of outer product calculation is too high. If the vector is of N dimension, the outer product is of n^2. Therefore, MCB proposes to map the result of the outer product to low dimensional space without explicit calculation of outer product.

4.6.1 Fusion strategy

In this experiment, to explore the influence of different image pre-training models and different fusion methods, different fusion strategies are used. In each fusion strategy, different image pre-training models are applied to extract image features. These image features and the text features of the output of Bert model are fused in the same fusion method. In detail, four different fusion methods are designed.

Concatenation: The image features are mapped to the same dimension as the text feature, and then two features are directly concatenated together as the fused features.

Element-wise Product Without Residual: The image feature is mapped to the same dimension as the text feature, then the image feature and the text feature are produced by the elements, and the final vector is taken as the fusion feature.

Table 8 The result of concate strategy

Image Extraction	Concat	Precision	Recall	F1	Accuracy
VGG16	Yes	0.7951	0.7395	0.7533	0.7911
ResNet50	Yes	0.8094	0.7877	0.7960	0.8182
DenseNet121	Yes	0.8315	0.7389	0.7564	0.8014

Element-wise Product With Residual: The image feature is mapped to the same dimension as the text feature, then the image feature and the text feature are product by elements, and the resulting vector and the original text feature are concatenated together as the fusion feature.

MurelBlock: MurelBlock [4] is the structure mentioned in an article of CVPR-2019, which is used for fusing image features and text features. In MurelBlock, the input image feature and text feature are firstly mapped to the same size dimension through bilinear mapping, then the two vectors are split according to the specified size. After that, the slices at the same location are further mapped with full connection layer without sharing parameters. Finally the results are concatenated as the fusion feature.

4.6.2 Experiment result

In this experiment, four sub-experiments are conducted. Every experiment use different fusion strategy to explore three image pre-training models. The following tables show the results of fused with text features.

From Table 8, the metric of concatenation strategy ResNet50 and DenseNet121 both better than the VGG16 Strategy, where ResNet50 is a little better. In aspect of image pre-training models, ResNet50 and DenseNet121 have better performance to get image features and the better features leads to a better result.

From Table 9, the conclusion is that whether use residual or not, the same image pre-training model has a similar result, which indicates that the residual does not work and not introduce more params.

From Table 10, the MurelBlock fusion strategy has the best result. It makes text features and image features better fused, and get a better representation to do fake news detection.

Table 9 The result of product residual strategy

Image Extraction	Residual	Precision	Recall	F1	Accuracy
VGG16	No	0.7951	0.7395	0.7533	0.7911
VGG16	Yes	0.7931	0.7424	0.7557	0.7917
ResNet50	No	0.8094	0.7877	0.7960	0.8182
ResNet50	Yes	0.8403	0.7781	0.7949	0.8259
DenseNet121	No	0.8315	0.7389	0.7564	0.8014
DenseNet121	Yes	0.8215	0.7627	0.7782	0.8117

Table 10 The result of MmurelBlock strategy

Image Extraction	MurelBlock	Precision	Recall	F1	Accuracy
VGG16	Yes	0.7890	0.7653	0.7739	0.7995
ResNet50	Yes	0.8374	0.8032	0.8152	0.8375
DenseNet121	Yes	0.8401	0.7642	0.7821	0.8182

Here are two conclusions as follows:

- To some extent, a better image pre-training model can get better image features. The appropriate image pre-training model should be chosen according to the actual needs.
- Traditional fusion strategies such as concatenation, element-wise product have similar performance, and the MurelBlock fusion strategy can get better result. Therefore, the Murelblock is used for feature fusion in later experiments.

4.7 Exp-IV: Strategy for statistic feature

The statistic feature refers to the numerical features obtained through statistics in the dataset. For fake news dataset, the user's relevant information can be regarded as statistical features, such as user followers, user fans, user location, etc. The experiments use different fusion strategies to test the influence of the fusion statistical feature. The specific fusion strategies are concatenation, element-wise product, and MurelBlock. The processing methods of statistical features include directly fusing as vectors, discretizing features, training different embedding, and fusing.

4.7.1 Fusion strategy

Through data preprocessing, the statistical features related to users are processed into a four-dimensional vector. The first dimension represents the gender of users, with values of 0, 1, or -1, the second dimension represents the number of user's Weibo fans, the third dimension represents the number of user's followers, and the fourth dimension represents the number of user's Weibo.

In terms of fusion strategy, we will use the MurelBlock to fuse text features and statistical features.

4.7.2 Experiment Result

In this experiment, the model is based on the baseline model. Table 11 shows the result.

Table 11 indicates that the statistical features improve the performance. The method of fusing statistical features is the same as that of fusing image features, so it is not elaborated too much here.

Table 11 The result of statistic feature fusion

Model	Precision	Recall	F1	Accuracy
Baseline	0.7897	0.7591	0.7693	0.8188
Baseline-Statistical	0.7821	0.7650	0.7717	0.7956

5 Conclusion

In this paper, we have conducted extensive experiments to investigate different approaches to multi-modal fake news detection. Combined with the experimental results, a new multi-modal architecture based on an attention mechanism and adversial mechanism is proposed to be designed, which performs best on fake news data. The existing multi-modal fusion strategies are still in the fine-tuning stage of the models. In the future, we will investigate multi-modal fusion strategies in the pre-training phase.

Declarations

Conflict of Interests Authors declare that they have no conflict of interest.

References

1. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D (2015) VQA: Visual Question Answering. International Journal of Computer Vision 123(1), 4
2. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
3. Bhatt G, Sharma A, Sharma S, Nagpal A, Raman B, Mittal A (2018) Combining Neural, Statistical and External Features for Fake News Stance Identification. In: Companion proceedings of the the web conference 2018, pp 1353–1357
4. Cadene R, Ben-Younes H, Cord M, Thome N (2019) MUREL: Multimodal Relational Reasoning for Visual Question Answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1989–1998
5. Cao J, Sheng Q, Qi P (2020) Progress and Prospects of Internet False Information Detection. Newsletter of the chinese computer society (3) 52
6. Cui L, Wang S, Lee D (2019) SAME: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, pp 41–48
7. Devlin J, Chang MW, Lee K, Toutanova K (2018) arXiv:1810.04805
8. Durier F, Vieira R, Garcia AC (2019) Can Machines Learn To Detect Fake News? A Survey Focused on Social Media
9. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding
10. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
11. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
12. Jin Z, Cao J, Zhang Y, Zhou J, Tian Q (2016) Novel visual and statistical image features for microblogs news verification. IEEE Transactions on Multimedia 19(3):598
13. Kafle K, Kanan C (2016) In: 2016 IEEE conference on computer vision and pattern recognition (CVPR)
14. Karpathy A, Fei-Fei L. (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
15. Ke L, Zou C, Bu S, Yun L, Gong M (2017) Multi-modal Feature Fusion for Geographic Image Annotation. Pattern recognition 73
16. Khattar D, Singh J, Gupta M, Varma V (2019) MVAE: Multimodal Variational Autoencoder for Fake News Detection. pp. 2915–2921. https://doi.org/10.1145/3308558.3313552
17. Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M (2016)
18. McCann B, Bradbury J, Xiong C, Socher R (2017) Learned in translation: Contextualized word vectors. In: Advances in neural information processing systems, pp 6294–6305
19. Mohtarami M, Baly R, Glass J, Nakov P, Màrquez L, Moschitti A (2018) arXiv:1804.07581
20. Parikh SB, Atrey PK (2018) Media-rich fake news detection: A survey. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (IEEE, 2018), pp 436–441

21. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
22. Qian F, Gong C, Sharma K, Liu Y (2018) In: Twenty-Seventh international joint conference on artificial intelligence IJCAI-18
23. Rashkin H, Choi E, Jin YJ, Volkova S, Choi Y. (2017) Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking
24. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
25. Shah P, Kobti Z (2020) Multimodal fake news detection using a Cultural Algorithm with situational and normative knowledge. In: 2020 IEEE Congress on Evolutionary Computation (CEC), pp 1–7. https://doi.org/10.1109/CEC48606.2020.9185643
26. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) ACM SIGKDD Explorations Newsletter 19(1):22
27. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556
28. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556
29. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Kabra A, Sharma M (2020) In: The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), pp 13,915–13,916
30. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S (2019) SpotFake: A Multi-modal Framework for Fake News Detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM), pp 39–47. https://doi.org/10.1109/BigMM.2019.00-44
31. Song C, Ning N, Zhang Y, Wu B (2020) A Multimodal Fake News Detection Model Based on Crossmodal Attention Residual and Multichannel Convolutional Neural Networks. Information Processing and Management 58. https://doi.org/10.1016/j.ipm.2020.102437
32. Sun C, Qiu X, Xu Y, Huang X (2019) In: China national conference on chinese computational linguistics (Springer), pp 194–206
33. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J. (2018) Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp 849–857
34. Wang Y, Qian S, Hu J, Fang Q, Xu C (2020) In: Proceedings of the 2020 International Conference on Multimedia Retrieval (Association for Computing Machinery, New York, NY, USA), ICMR '20, p 540–547. https://doi.org/10.1145/3372278.3390713
35. Zhang Q, Yilmaz E, Liang S (2018) Ranking-based method for news stance detection. In: Companion proceedings of the the web conference 2018, pp 41–42
36. Zhou X, Wu J, Reza Z (2020) SAFE: Similarity-Aware Multi-Modal Fake News Detection. arXiv:2003.04981