# Email False Information Detection Based on Attention Mechanism and Contrastive Learning

2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 978-1-6654-9125-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIBD57115.2023.10206391

Jun Li*
*Beijing Information Science and Technology University*
Beijing, China
lijun@bistu.edu.cn

Zhuoyuan Li
*Beijing Information Science and Technology University*
Beijing, China
492720572@qq.com

Yanzhao Liu
*China Information Technology Security Evaluation Center*
Beijing, China
liuyz@itsec.gov.cn

Shuo Wang
*Dalian University of Technology*
Dalian, China
ws13838238055@mail.dlut.edu.cn

***Abstract*—Email has become an important way for people to communicate with each other. Because of its high security and fast sending and receiving characteristics, it can establish a corporate brand image while being portable and managed. key tool. Therefore, many criminals take advantage of the feature of mailboxes to send false information through mailboxes. This kind of false information mostly appears in the form of cross-modality, which is very confusing and can easily cause extremely bad effects. To solve this problem, this paper proposes a multi-modal false information detection method using contrastive learning pre-training and attention mechanism, uses contrastive learning to align features between different modal data, and uses attention mechanism to achieve different modal features. The interaction between them, the model construction is completed through feature fusion, and finally the accurate detection of fake emails is realized, so as to facilitate the next step of traceability and countermeasures. Compared with the current mainstream methods, the model proposed in this paper has achieved better results in the detection of multi-modal false information, with an increase of more than 6% in accuracy and other aspects, and can achieve more accurate identification and detection of cross-modal false emails.**

***Keywords—false information detection, machine learning, computer vision, natural language processing(NLP), feature extraction***

## I. INTRODUCTION

With the development of the network and the continuous upgrading and popularization of computer hardware equipment, the way of communication between people has gradually developed from writing paper letters to sending emails now. While emails bring convenience to people, they also bring A lot of questions came up. Bearing the brunt of the problem is the emails include false information, which has caused bad influence and huge financial losses. And because it mostly appears in the form of cross-modality, it is extremely difficult to detect, so it is very urgent to solve such problems.

The early method to solve this kind of problem is to extract language features from the text for single-modal mail false information detection. Later, with the deepening of the research of convolutional neural network, it is also proposed to extract the features from the text and from the picture. The extracted features are combined to detect multi-modal false information. These methods take into account the relationship between different modal data features, and realize the detection of multi-modal false information by interacting the relationship between different modalities. Although these schemes are feasible to some extent, there are also some shortcomings. First, the eigenvectors of different modalities are in different eigenspaces, and it is difficult for the eigenvectors of different modalities to form a better interaction relationship. Secondly, in other multimodal joint representations, many of them ignore the interaction relationship within a single modality and only focus on the interaction between different modalities. In order to solve the above problems, this paper proposes a multimodal attention network model based on contrastive learning pre-training:

(1) In terms of data preprocessing, the BERT[1] pre-training model is used to extract the language features of the text modal data in the sample to be tested, and the ResNet[2] network is used to extract the visual features of the image modal data in the sample to be tested, and then the The feature vectors extracted from different modal data are pre-trained by contrastive learning[3], and the feature vectors of different modalities are mapped to the same feature space by minimizing the contrastive loss function to achieve feature alignment.

(2) A cross-modal attention mechanism[4] is introduced to capture the high-level interaction between different modalities, and the feature vectors of visual and language modalities are updated according to the correlation weights learned between different modalities. At the same time, the complex relationship between different feature vectors in a single modality is obtained through the intra-modal self-attention mechanism.

(3) The final joint representation is obtained by fusing the features of different modalities, and the final joint representation is projected into the binary classification space[5] to realize the authenticity judgment of the information to be tested.

## II. RELATED WORKS

For the problem of false email detection, we can define it as the problem of false information detection. At this stage, the detection of false information and rumors has become a research hotspot at home and abroad, mainly using machine

**Foundation item:** National Natural Science Foundation of China (Grant No. U1936111) && Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University && Equipment Development Department's Insight Action Project(F2B6A194). Correspondence should be addressed to Jun Li: lijun@bistu.edu.cn

learning and deep learning methods,and has achieved effective research results.

In terms of fake email detection, the mainstream method is to obtain the content features of different modalities in the email, and form a joint representation to detect fakeness. Liu Jinshuo et al [6] realized the detection of false information on the network by constructing the MSRD model, extracted the features of the picture through the VGG19 network, then extracted the embedded text in the picture through DenseNet, and used the LSTM network to extract the features of the embedded text. The language features of the text and the visual features of the pictures are spliced together, and the mean and variance vectors of the shared representation of the language and visual levels are obtained through the fully connected layer, and finally the matching degree detection of the embedded text and pictures is realized, and the detection of false information is realized. screening. Khattar et al. [7] reconstructed the data through an autoencoder to obtain a shared representation, and then performed binary classification on it to achieve authenticity identification. The above two methods simply achieve joint representation through feature splicing, which ignores the relationship between different modal features, and the accuracy is poor when performing downstream detection tasks. Gao et al. [8] established a multi-modal feature fusion dynamic fusion internal model and inter-mode attention flow model. This method is also mentioned in the article of Meng et al. [9]. Through Faster-RCNN and gated recursive unit extraction The feature vectors of text data and image data, and interact with the feature vectors of different modalities by establishing the intra-modal and inter-modal attention flow framework (DFAF), and multiple times for the intra-modal and inter-modal attention flow framework Iteration, deep fusion of different modal eigenvectors to obtain a joint to detect false information, this method uses the attention mechanism to capture the relationship between different modalities, but the eigenvectors of different modalities belong to different feature spaces , the detection effect still needs to be improved. WenhuiWang et al. [10] proposed the BeiT-v3 model. By using a pre-training task to obtain a general multimodal model, this model does not need to consider different inputs to be processed differently. The author directly processes images and texts in the same way. Alignment modes, which also provide new ideas for our research.

## III. Problem Description

False information detection on platforms such as social media is essentially a binary classification task for multimodal events. The given event group is $U=\{u_1;u_2;u_3;...u_n\}$, the U collection consists of n elements, and each element in the collection represents a fixed event, such as representing the i-th event, which contains the text related to the event class data and corresponding image class data. Given the corresponding label group $L=\{l_1;l_2;...l_n\}$, each element in the label group represents the label of the corresponding event, and these labels mark the corresponding event, and are divided into false information labels and non-false information labels.

In the process of false information detection, the event group U set is first divided into training set $U_{train}$ and test set $U_{test}$. Input the collection $U_{train}$ and the collection $U_{test}$ into the model for learning, so as to build a mapping model with event $u_i$ as input data and label $l_i$ as output data. Then the test set is used as the input of the model, and the corresponding predicted label set $L_{pre}$ is output. Finally, the matching probability is obtained through the fully connected layer to judge the detection effect of the model on emails with false information.

## IV. Model Introduction

The multimodal attention detection network (MADN) is mainly used for feature extraction, feature fusion, feature matching, and the function of generating authenticity probability for multimodal information to determine whether it is false information.

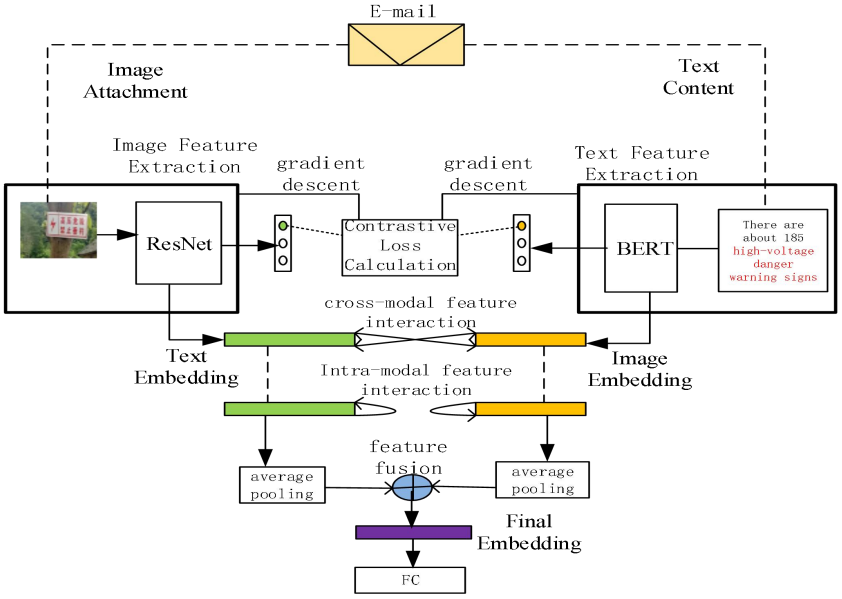


Fig. 1. Multimodal attention detection network.

Its main structure is shown in Figure 1. The multimodal information to be detected mainly consists of text language data and graphic and visual data. In order to realize the key element extraction of multimodal data such as text and pictures, a multimodal attention mechanism is established. First, the BERT pre-training model is used to extract the features of the text data, and then the ResNet network is used to extract the features of the image modality data, and then the feature vectors of the two modalities are compared and pre-trained. Finally, the multi-modal attention mechanism is used to capture the high-level interaction between the language and visual fields and realize the final representation of multi-modality. The final representation is classified through the fully connected layer to obtain the authenticity probability of multi-modal information. The multimodal attention mechanism structure is mainly composed of four parts: text feature extraction, image feature extraction, feature fusion mechanism and information classifier.

### *A. Text Feature Extraction Module*

Natural language processing (NLP) is divided into upstream tasks and downstream tasks. The upstream task is responsible for pre-training the data. In the current related scheme design, models with high frequency of use in upstream tasks include ELMo[11], Word2Vec[12], etc. Among them, the core idea of the ELMo model is embodied in the deep context relationship. The context-dependent word vector representation is obtained through the bidirectional language model according to the specific input, which enhances the vector feature representation ability, but the work efficiency is low. The Word2Vec model will map the words in a sentence into word vectors through a high-dimensional space, which can represent the relationship between words. Each word processed by Word2Vec corresponds to a fixed word vector, but it does not consider the word order problem, and cannot consider the context correlation in the entire long sentence, and cannot understand the semantics of the context.

The BERT model obtains the expression of language information by training large-scale unlabeled corpus data. The model has the characteristics of good parallelism and high efficiency, and can convert data into sentence vectors with consistent dimensions. Therefore, this paper uses the BERT model to pre-train language modality data to obtain language modality feature vectors. The length of the corresponding text is truncated and filled to 15, expressed as T=[$T_0, T_1, T_2 \dots T_{15}$], $T_0$ is expressed as [CLS] embedding, and then input to the BERT model to extract the word feature vector as $T \in \mathbb{R}^{15*2048}$.

### *B. Image Feature Extraction Module*

Compared with the traditional VGG network, the ResNet network has better performance in terms of feature extraction and target classification. The characteristic of the residual network is that it is easy to optimize and can improve the accuracy by adding considerable depth. Its internal residual block uses skip connections, which alleviates the problem of gradient disappearance caused by increasing depth in deep neural networks. Especially when the network level is deeper, the ResNet network solves the network degradation problem that occurs in the traditional VGG network. The image feature extractor is responsible for capturing image features. Given that the input data $u_i$ is the data in the set U, which is the picture-text data pair to be detected, the basic feature extraction is performed on the picture data of the event in the training set through the ResNet50 network, and finally the visual feature is obtained as $I \in \mathbb{R}^{2048}$.

### *C. Feature Fusion Mechanism*

The feature fusion mechanism is mainly composed of five parts: contrastive learning pre-training, inter-modal attention interaction mechanism, intra-modal attention interaction mechanism, multi-modal fusion and Information detection.

#### *1) Contrastive learning pre-training*

Before the interaction of different modal features, the image-text contrast learning loss function (Image-Text Contrast Loss, ITC)[13] is used to align the features of images and texts. After text feature extraction and picture feature extraction, the text feature vector and picture feature vector are obtained. Afterwards, the similarity score is obtained by learning the similarity function, and the similarity function is as follows:

$$\boldsymbol{s = g_v(v_{item})^{\intercal} g_w(w_{item})} \tag{1}$$

Among them, $g_v$ and $g_w$ are linear representations that map feature vectors to normalized low-dimensional spaces. Then calculate the similarity between the graphics and text, the calculation formula is as follows, where the formula (2) is the similarity between the visual mode and the language mode feature, and the formula (3) is the similarity between the language mode and the visual mode feature:

$$\boldsymbol{y_i(I) = \frac{e^{(s(I,T_i)/\tau)}}{\sum_{i=1}^{M} e^{(s(I,T_i)/\tau)}}} \tag{2}$$

$$\boldsymbol{y_i(T) = \frac{e^{s(T,I_i)/\tau}}{\sum_{i=1}^{M} e^{s(T,I_m)/\tau}}} \tag{3}$$

Among them, $\tau$ is a learnable temperature parameter, $y_i(I)$ and $y_i(T)$ represent the one-hot similarity of ground truth, the probability of a negative sample pair is 0, and the probability of a positive sample pair is 1. The cross-entropy loss of graphic-text contrastive learning is defined as follows:

$$\boldsymbol{\mathcal{L} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}[H(y(I), y(I)) + H(y(T), y(T))]} \tag{4}$$

In this model, the most commonly used InfoNCE loss function[14] when calculating contrastive learning loss is adopted. During the training process, the features of the image and text positive sample pairs are narrowed by minimizing the InfoNCE loss function value. The specific function form is as follows:

$$\boldsymbol{\mathcal{L}_{I\sim T} = -\frac{1}{2}\mathbb{E}_{(I,T)}[log \frac{e^{(s(I,T)/\tau)}}{\sum_{i=1}^{M} e^{(s(I,T_i)/\tau)}}]} \tag{5}$$

$$\boldsymbol{\mathcal{L}_{T\sim I} = -\frac{1}{2}\mathbb{E}_{(T,I)}[log \frac{e^{(s(T,I)/\tau)}}{\sum_{i=1}^{M} e^{(s(T,I_i)/\tau)}}]} \tag{6}$$

$$\mathcal{L} = \mathcal{L}_{I\sim T} + \mathcal{L}_{T\sim I} \quad (7)$$

Among them, formula (5) is the calculation process of the InfoNCE loss function between the visual mode and the language mode, formula (6) is the calculation process of the InfoNCE loss function between the language mode and the visual mode, and $\mathcal{L}$ is the graphic-text comparison learning The value of the loss function.

Then, gradient descent optimization is performed according to the loss function $\mathcal{L}$ to achieve feature alignment of features in two independent feature spaces, thereby mapping to the same feature space, enhancing the feature correlation between vectors, and preparing for the subsequent cross-modal feature interaction.

*2) Inter-modal attention mechanism*

The intermodal attention module first uses the attention mechanism to calculate the correlation between different modalities, so as to update the feature vectors of text and pictures through the learned correlation weights. The calculation formula of inter-modal attention is as follows.

$$\boldsymbol{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \boldsymbol{softmax}(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d}}})\boldsymbol{V} \quad (8)$$

Among them, is the attention operation function, Q, K, V are the query matrix, key matrix and value matrix respectively; d is used as a scaling factor to prevent the molecular dot product value from being too large, and its value is the dimension of the input feature.

Input the text feature vector Text Embeding (represented by T below, $T \in \mathbb{R}^{15*2048}$) and the image feature vector Image Embeding (represented by I below, $I \in \mathbb{R}^{2048}$) into the intermodal attention mechanism module. The attention operation function process is as follows.

$$\boldsymbol{T_{update}} = \boldsymbol{Attention}(\boldsymbol{Q_T},\boldsymbol{K_I},\boldsymbol{V_I}) \quad (9)$$

$$\boldsymbol{I_{update}} = \boldsymbol{Attention}(\boldsymbol{Q_I},\boldsymbol{K_T},\boldsymbol{V_T}) \quad (10)$$

After calculation, the updated text feature matrix $T_{update}$ and image feature matrix $I_{update}$ are obtained as the input matrix of the intra-modal attention mechanism.

*3) Intra-modal attention mechanism*

We model single intra-modal relations using an intra-modal attention module, which is computed as follows.

$$I_{new} = Attention(Q_I, K_I, V_I) \quad (11)$$

$$T_{new} = Attention(Q_T, K_T, V_T) \quad (12)$$

We can get the updated text feature $T_{new}$ and image feature $I_{new}$.

*4) Multi-modal Fusion*

After the above process, average pooling will be performed on the obtained features, and finally the final representations of different modes are obtained, which are $T_f$ and $I_f$ respectively.

$$T_f = AvgPool(T_{new}) \quad (13)$$

$$I_f = AvgPool(I_{new}) \quad (14)$$

Finally, the final representation sum after pooling is spliced and linearly transformed to obtain the final joint representation Final Embedding.

*5) Information detection*

This module is based on the final joint representation Final Embedding obtained above, and uses the fully connected layer (FC) with the activation function softmax to project the final representation Final Embedding to the binary classification target space, thereby obtaining the probability distribution p. The softmax operation function process is as follows.

$$\boldsymbol{p_i} = \boldsymbol{softmax}(\boldsymbol{u_i}) = \frac{\boldsymbol{e^{u_i}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e^{u_i}}} \quad (15)$$

Among them, is the i-th input data $u_i$ in the event group U, which is subjected to binary classification through the softmax function to obtain its authenticity probability $P_i$.

## V. Experiment

### A. Experimental Dataset

In order to evaluate the multimodal attention detection network model, the data set used in this experiment is obtained by data enhancement after obtaining the data of a company's work mailboxes (mainly 163 mailboxes and QQ mailboxes). The data set contains 7608 pairs of true information and 7085 pairs of false information. The statistics of the dataset are shown in Table 1.

TABLE I. Statistics of the Dataset

| platform | True | False | numbers of picture |
|---|---|---|---|
| 163 mailbox | 3876 | 3659 | 7535 |
| QQ mailboxes | 3732 | 3426 | 7158 |


Fig. 2. Examples of Dataset.

### B. Experiment setup

The hardware and software environment is: AMD Ryzen7 5800H 3.2GHZ, RTX-3090 GPU, Python3.8.12. The experimental configuration and parameters are shown in Table 2:

TABLE II. THE EXPERIMENTAL CONFIGURATION AND PARAMETERS

| parameter | value |
|---|---|
| batch-size | 32 |
| lr | 0.001 |
| dropout | 0.3 |
| epoch | 50 |
| optimization | Adam |
| GPU | RTX3090 |

### C. Evaluation Method

In this experiment, the ResNet network and BERT are used to extract image and text features, and then the attention mechanism is used to fuse the features of different modalities to detect false information mixed with different modalities.

Commonly used evaluation indicators used in this experiment are: Accuracy, Precision, and Recall. TP indicates the number of image-text pairs that are correctly identified as rumors, and FP indicates the number of image-text pairs that are incorrectly identified.

$$Accuracy = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$Precision = \frac{TP+TN}{TP+TN+FP+FN}$$

### D. Comparative Experiment

The experimental results of the benchmark model and the MADN model on the mailbox dataset are shown in Table.3. The experimental results show that the MADN model in this paper is superior to the benchmark model in terms of accuracy, F1 and other indicators.

After using contrastive learning pre-training to pre-train the data of visual modality and language modality, the feature alignment between different modalities is realized, so that the interaction between feature vectors of different modal data can be better realized. After attention interaction between different modalities and within the same modality and feature vector fusion, the detection of multi-modal false information will have a better performance.

1) MMDF[9]: The MMDF model extracts different modal feature vectors through multi-level RNN-CNN and Bi-GRU respectively. A cross-modal self-attention mechanism and a cross-modal co-attention mechanism are established, and an intermodal attention flow is established through the attention mechanism, which realizes feature extraction and feature fusion of data of different modalities.

2) MVAE[7]: The model detects disinformation by learning shared representations of both textual and visual modality data.

3) EANN[15]: The EANN model forms multi-modal features by splicing visual features and visual features. Finally, the detection of false information is realized through the time classifier.

TABLE III. COMPARISON EXPERIMENT RESULTS

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| MMDF | 0. 721 | 0.706 | 0.711 | 0.708 |
| MVAE | 0. 764 | 0.735 | 0.735 | 0.758 |
| EANN | 0. 753 | 0.801 | 0.801 | 0.810 |
| MADN(ours) | **0. 857** | **0. 851** | **0. 836** | **0. 843** |

### E. Ablation Experiment

In order to understand the role of each module of the MADN model more clearly, we split each module of the model structure in an incremental form, and gradually assembled it to conduct comparative experiments from simple to complex. The results of the ablation experiment are shown in Table 4.

1) MADN: Contains all modules of the MADN model.

2) w/o inter-att/intra-att: It consists of text feature extraction part, image feature extraction part and contrastive learning pre-training part.

3) w/o contrastive learning: It consists of a text feature extraction part, an image feature extraction part, and an inter-model/intra-model attention interaction part.

4) Text+Image: It consists of a text feature extraction part and an image feature extraction part. The visual feature vector and language feature vector are obtained, and then average pooled and spliced into the final joint representation.

TABLE IV. ABLATION EXPERIMENT RESULTS

| Model | Accuracy |
|---|---|
| MADN | 0.857 |
| w/o contrastive learning | 0.801 |
| w/o inter-att/intra-att | 0.771 |
| Text+Image | 0.623 |

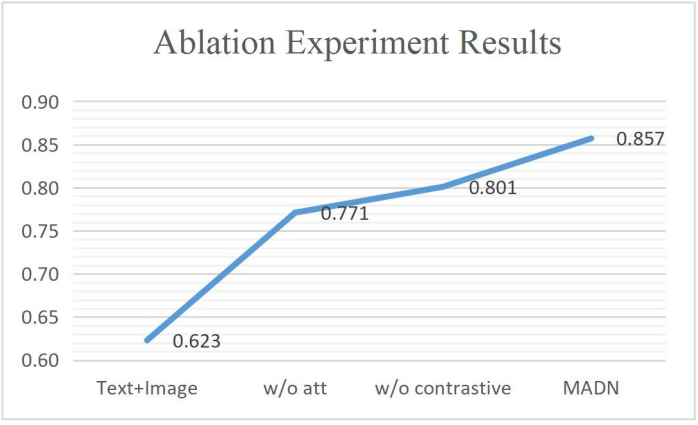


Fig. 3. Line chart of ablation experiment results

From the data results of the ablation experiment and the line graph, it can be seen intuitively that in the absence of contrastive learning pre-training and attention mechanism, The multimodal model detecting cross-modal fake emails that only combines features of different modal data is extremely terrible. When the cross-modal attention mechanism is added after the feature extraction of different modalities, the detection accuracy of this model is improved by about 15-18 percentage points, which proves that inter-modal attention and intra-modal attention can effectively realize different modes. The interaction between modalities can improves the detection effect of cross-modal false information in emails. When only contrastive learning is used to align features of different modalities without attention interaction between modalities, the detection effect is not good, but the accuracy is still greatly improved compared with the simple feature splicing model. On this basis, the contrastive learning pre-training of different modal features is carried out, and the accuracy rate of the obtained MADN model for false information identification is about 86% after the inter-modal and intra-modal attention interaction, compared with the previous model in the email The accuracy of false information detection has been greatly improved, and the accuracy has increased by more than 6%. Facts have proved that after feature alignment of different modal features through contrastive learning pre-training, and cross-modal interaction through attention mechanism, the accuracy of false information detection can be significantly improved.

## VI. CONCLUSIONS

In the process of realizing the false information detection task of email, after performing feature alignment on different modal data through comparative learning pre-training, and then through the inter-modal and intra-modal attention mechanism, it is possible to better realize the difference between different modalities. The interaction among them, especially after the pre-training of each modal data through comparative learning, strengthens the association between weakly correlated positive sample pairs, and greatly improves the accuracy of the model for email false information detection. A new solution is provided for the detection task of fake emails.

## REFERENCES


[1] Devlin Jacob,Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019, June 2, 2019.

[2] He Kaiming, Zhang Xiangyu and Ren Shaoqing and Sun Jian. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,December,2016. 770 - 778.

[3] He Kaiming, Fan Haoqi, Wu Yuxin, Xie Saining and Girshick Ross. Momentum Contrast for Unsupervised Visual Representation Learning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, 9726 - 9735.

[4] De Santana, Correia Alana, Colombini, Esther Luna. Neural attention models in deep learning: Survey and taxonomy. arXiv. 2021.

[5] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. SpotFake：a multi-modal framework for fake news detection［C］// Proceedings of the IEEE 5th International Conference on Multimedia Big Data. Piscataway：IEEE，2019：39-47.

[6] LIU Jinshuo, FENG Kuo, Jeff Z. Pan, DENG Juan, WANG Lina. MSRD: Multi-Modal Web Rumor Detection Method[J]. Journal of Computer Research and Development, 2020, 57(11): 2328-2336.

[7] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In The World Wide Web Conference (WWW '19). Association for Computing Machinery,NewYork,NY,USA,2915–2921.

[8] GAO Peng, JIANG Zhengkai., YOU Haoxuan, LU Pan, HOI S. C. H., WANG Xiaogang, & LI Hongsheng. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]//IEEE.32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16, 2019，CA, United states. New York：IEEE, 2019：6632–6641.

[9] MENG Jie, WANG Li , YANG Yanjie , LIAN Biao. Multi-modal deep fusion for false information detection[J]. Journal of Computer Applications, 2022, 42(2): 419-425.

[10] Wang Wenhui, Bao Hangbo, Dong Li, et al. Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks. arXiv, 2022.

[11] Peters Matthew E, Neumann Mark, Iyyer Mohit, Gardner Matt and Clark Christopher. Deep contextualized word representations. arXiv, 2018.

[12] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey. Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013.

[13] Hou M, Xu C, Liu Y, et al. Stock Trend Prediction with Multi-granularity Data: A Contrastive Learning Approach with Adaptive Fusion[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 700-709.

[14] Yang, Ji and Yi, Xinyang and Zhiyuan Cheng, Derek and Hong, Lichan and Li, Yang and Xiaoming Wang, Simon and Xu, Taibai and Chi, Ed H. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020,2020,441-447.

[15] WANG Yaqian, MA Fenglong, JIN Zhiwei, YUAN Ye, XUN Guangxu, JHA K, SU Lu, & GAN Jing. EANN: Event adversarial neural networks for multi-modal fake news detection[C]//ACM. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2018, August 19, United kingdom. London:ACM, 2018, 849–857.