Information Fusion 93 (2023) 150–158

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

TRIMOON: Two-Round Inconsistency-based Multi-modal fusion Network for fake news detection

Shufeng Xiong [a], Guipei Zhang [a], Vishwash Batra [b], Lei Xi [a], Lei Shi [a], Liangliang Liu [a,*]

[a] Henan Agricultural University, Zhengzhou 450002, China
[b] School of Computer Science and Mathematics, Keele University, Keele ST5 5AA, UK

ARTICLE INFO

Keywords:
Fake news detection
Multi-modal fusion
Deep learning
Feature fusion

ABSTRACT

Compared to ordinary news, fake news is characterized by faster dissemination and lower production cost and therefore causes a great social harm. For these reasons, the challenge to efficiently and accurately detect fake news has attracted a lot of attention in the research community. We propose a **T**wo-**R**ound **I**nconsistency-based **M**ulti-m**o**dal fusi**o**n **N**etwork (TRIMOON) for fake news detection, which consists of three main components: the multi-modal feature extraction module, the multi-modal feature fusion module and the classification module. To filter the noise generated in the fusion process, we perform a two-fold inconsistency detection, once before and once after the fusion process. Experimental results also prove this to be quite effective. Our proposed TRIMOON is evaluated on both the Chinese and the English datasets, and our model outperforms the state-of-the-art approaches on several classification evaluation metrics.

1. Introduction

Internet and social media have surely made human life convenient, but it has brought some adverse effects too. Fake news is one such example. With the advent of social media, the spread of fake news has rapidly increased and has caused detrimental effects on society. False news refers to deliberately providing people with incorrect information with the intention of misleading [1–3]. False news makers usually release false information that is easy to attract people's attention to obtain political or monetary benefits and use automatic methods to spread information at a faster pace. They also employ various other strategies to make it challenging for fake news to be detected [4].

In recent years, the effectiveness and therefore the negative impact of fake news has significantly increased, greatly influencing public opinion. For example, in the wake of the Corona Virus Disease 2019 (COVID-19) outbreak, fake news has proliferated around the world, causing social and mainstream media to be flooded with rumours [5], adversely affecting measures for epidemic control. In addition, studies have shown that fake news was an important factor in Donald Trump's victory in the 2016 election [6]. Fake news also has financial implications. In 2013, a fake message mentioned that Barack Obama was injured in a White House bombing. The tweet sent the S&P 500 down 0.9%, according to the Financial Times [7].

Considering the great harm caused by fake news, it is imperative to come up with appropriate solutions. The researchers initially used traditional machine learning methods to detect fake news, such as Support Vector Machine (SVM) and Naive Bayes Classifier, which differ in function and structure but produce similar results and are used as baseline models [8,9]. In addition, clustering algorithms and Decision Trees have widely been used in experiments [10]. With the emergence of Deep Learning, these models have become mainstream at various tasks. Sastrawan et al. [11] use Convolutional Neural Network joined by a Recurrent Neural Network (CNN–RNN) to detect false news. Choudhary et al. [12] proposed BerConvoNet framework, which combined Bidirectional Encoder Representation from Transformers (BERT) Embedding and CNN for detecting false news. Rai et al. [13] proposed a false news classification method based on news headlines by combining BERT and Long Short-Term Memory (LSTM) network.

Although deep learning models have achieved good results, most studies only consider text features ignoring image features. Therefore, some scholars have proposed multi-modal based detection methods. Kumari et al. [14] proposed a multi-modal framework based on a deep neural network, Attention-based multi-modal Factorized Bilinear Pooling (AMFB), which combines text and image features together and detects fake news through a Multilayer Perceptron (MLP) model with two hidden layers and an output layer with Sigmoid activation. Song et al. [15] proposed a multi-modal fake news detection framework based on Crossmodal Attention Residual and Multichannel convolutional neural Networks (CARMN). The Crossmodal Attention Residual Network (CARN) not only integrates the relevant information between

* Corresponding author.
E-mail address: liangliu@henau.edu.cn (L. Liu).

https://doi.org/10.1016/j.inffus.2022.12.016
Received 8 October 2022; Received in revised form 9 December 2022; Accepted 16 December 2022
Available online 30 December 2022

Fig. 1. News with inconsistent text and pictures.

different modalities but also maintains the unique attributes of each modality. The Multichannel Convolutional neural Network (MCN) can extract feature representations from the original and fused text information at the same time. This model is able to learn more discriminable feature representations and thus perform better at various tasks.

However, sometimes there is inconsistency between texts and images of a news article. For example, in Fig. 1, the text of the news from Hollywood Life (A website that reports the latest Hollywood gossip, news and celebrity photos) is about Jared Leto and Angelina Jolie having a secret date, but the picture is made up of two pictures, so people cannot get any useful information from this. The next day, Gossip Cop (A website that covers entertainment news and clears up rumours) published a debunking story, pointing out that it was a complete fabrication. Therefore, to judge the consistency of images and texts, Xue et al. [16] proposed a multi-modal consistency neural network (MCNN), which contains a similarity measurement module, which uses cosine distance to measure the consistency between images and texts. This experiment takes into account the consistency of multi-modal data, which greatly enhances detection of such false news+ cases where images do not match. In terms of multi-modal based false news detection, this model is effective than other baseline models.

In summary, although the existing research has focused on the two methods of multi-modal fusion of images and texts and image-text consistency detection and have also achieved good performance, but the current methods still face the following challenges: (1) Multi-modal feature fusion and image-text consistency discrimination are performed independently, which leads to presence of noise in the final fusion features. (2) As text is the dominant part of a news article, in existing models, information in textual modality is not yet fully reflected in its leading role in expressing information. Therefore, the challenge to fully consider the fusion of image and text features, highlighting the dominant position of text in news, and also taking into account the issue of image-text consistency simultaneously is under-explored.

To address the above challenges, we proposes a multi-modal fusion model, called a two-round inconsistency-based multi-modal fusion network (TRIMOON), that takes into account image-text inconsistency. First, we built a semantic consistency scoring module preceding the general image-text co-attention fusion module, whose purpose is to control the fusion strength based on the degree of consistency. Secondly, we perform another fusion of text modality and image-text fusion representation, to strengthen the dominance of the text information. Finally, based on the fusion representation, the Bi-directional Long Short-Term Memory (BiLSTM) is used to encode the document representation and output the classification labels through the fully connected layer.

The main contributions of this paper are as follows.

- We propose a multi-modal fusion module for fake news detection based on the degree of image-text consistency, which effectively suppresses the noise generated by the fusion representation when the images and texts are inconsistent.
- We provide a secondary fusion mechanism for the representation of the fusion of image-text information, thereby strengthening the leading role of text modalities in news media.
- We perform experimental comparisons with the State-Of-The-Art (SOTA) method on publicly available Chinese and English datasets to verify the effectiveness of the proposed model.

The subsequent sections of this paper are organized as follows. Section 2 provides an analysis of related work and summarizes the differences between our proposed approach and existing methods. Section 3 provides a detailed description of the TRIMOON model proposed in this paper. Section 4 provides the setup of the verification experiments and the observation and analysis of the experimental results. Section 5 provides discussions on some of the issues in modelling that need addressing. The last section presents concluding remarks indicating future research directions alongside.

2. Related work

As discussed in the last section, supervised learning methods are the mainstream for dealing with fake news detection, although there are currently some unsupervised methods [17] or social network-based methods [18]. The mainstream methods fall in two categories: (1) methods based on traditional machine learning and (2) methods based on deep learning.

2.1. Based on traditional machine learning

Automatic detection of false news is at an early stage to become the mainstream method. As the volume of digital news is increasing rapidly, manual detection of fake news requires extensive human labour. Researchers have developed fake news detection methods based on traditional machine learning.

Castillo et al. [19] from the perspective of users' communication behaviour, constructed feature sets such as content, topic, communication and behaviour, and verified them by SVM [20], J48 [21] and other machine learning algorithms, finally reaching 89% classification accuracy. The experiment verified the effectiveness of introducing the above four types of features. Ruchansky et al. [22] proposed an automatic fake news detector called CSI (Capture, Score, and Integrate). It consists of three modules: capture, scoring, and integration. The detector predicts fake news by using three features associated with incoming news: text, response, and source. The model consists of three modules. The first one extracts the time representation of the news article. The second represents the user's behaviour and scores it, and the final module uses the output of the first two modules and uses them for categorization. Their experiments show that CSI has improved in accuracy. Reis et al. [23] proposed a new feature set for this task, and apply it to the existing automatic detection model, and experimentally verify the effectiveness of the new feature set. Furthermore, one of the aims of fake news is to manipulate people's opinions, so sentiment information [24] is also considered as a feature in some work [25,26].

Explicit features are easy to interpret, but not necessarily the best data to learn for machine learning models. Lately some research works have started exploring implicit features of fake news. Guo et al. [27] constructed features based on user portraits according to news data, incorporating implicit features such as user behaviour, credibility and reliability. Wu et al. [28] proposed that news topic types and emotional features, combined with message propagation interval features, and the support vector machine algorithm with Random Walk [29] kernel function was used for classification, which achieved good results in micro-blogging data sets.

2.2. Based on deep learning and multi-modal

Fake news detection methods based on traditional machine learning greatly improve the efficiency of detection of fake news. However, as news content is becoming more complex, reliance on artificial features alone will not be enough. Also, with the popularity of deep learning algorithms, many researchers have started employing deep neural network based models for the task of fake news detection [12,13,30,31]. Relying on the ability of automatic feature extraction, they can dig out features that are more essential and easier to learn than artificial features used in traditional machine learning. Meanwhile, multi-modal information is also introduced into fake news detection through deep neural networks.

Jin et al. [32] proposed an attention-based Recurrent Neural Network (att-RNN) that incorporates information from text, visual and social contexts. Wang et al. [33] proposed an Event Adversarial Neural Network (EANN), which introduces event classification tasks into adversarial learning and guides the model to learn event-independent text modal and image modal features with more generalization performance. Khatter et al. [34] proposed an End-to-End multi-modal Variational Autoencoder (MVAE) network, which consists of three main components, an encoder, a decoder, and a fake news detector Module, which used the encoder–decoder structure to construct the feature expression of multi-modal news.

The above methods are effective in detecting fake news with multi-modal information, but they cannot fully understand the deep semantics of multi-modal news events due to the lack of sufficient factual knowledge. To solve this problem, Zhang et al. [35] proposed a novel Multi-modal knowledge-aware Event Memory Network (MKEMN), which utilizes multi-modal Knowledge awareness Network (MKN) and Event Memory Network (EMN) as building blocks of social media rumour detection. The conceptual knowledge corresponding to text entities is extracted from the external knowledge map and integrated into the multi-modal representation to obtain higher semantic understanding capability. Wang et al. [36] propose a novel Knowledge-driven Multi-modal Graph Convolutional Network (KMGCN), which models semantic representation through joint modelling of text information and integrates Knowledge concepts and visual information into a unified fake news detection framework. Indeed, the availability of images is also an issue to be considered in multi-modal fusion models. Some work considers the alignment of text with the entities appearing in the image to distinguish inconsistencies between the image and the text to enhance detection performance [37,38]. Subsequently, Li et al. proposed an entity-centric multi-modal learning framework with an approach that learns new feature representations to train classifiers through two modules: entity alignment and entity-centric feature aggregation [39]. Chen et al. on the other hand, proposed a cross-modal ambiguity learning model for dynamic fusion of uni-modal and multi-modal features, with uni-modal features dominating when the ambiguity is large and cross-modal features dominating when the ambiguity is small [40].

We propose a new multi-modal fusion approach from the perspective of image and text consistency. Unlike entity-level consistency [37,38], we consider the overall semantic consistency of images with the news text [16,40]. In our scheme, we highlight text modality as the dominant information of news while appropriately fusing image modality information based on inconsistency, while [16] aligns consistency directly with class labels, and [40] learns the consistency representation of image and text modalities from the dataset overall level and performs weighted fusion based on this. In our model, whether the image information is consistent with the text content or not, it does not replace the text modal features as the dominant feature for classification, which is also in line with our human habit of reading and judging the reality or falsity of news.

3. Methodology

The previous paragraph has analysed how our approach differs from existing methods. In this section, the specifics of the proposed TRIMOON model are described in detail. In order to better understand the idea of our proposed method, we begin with formal problem description.

Given a collection $D = \{(x_i, y_i), i = 1, \ldots, N\}$ of N news items, each news x_i has a label $y_i \in \{0, 1\}$ indicating its category, x_i contains x_i^T and x_i^V, where x_i^T is the text in the news item, x_i^V is the image in the news item. Here, each entry of the label vector y_i of a sample x_i indicates whether it belongs to a certain class (1) or not (0). Our proposed model will utilize training instances to learn a mapping function $F : X \rightarrow Y$ from the feature space X to the label space Y, and then use F to predict the label vector of an instance based on its feature vector. The input of our model is the text and image of each news item. BERT and VGG (Visual Geometry Group) networks are used to learn the text features and image features respectively.

In our work, we believe that when detecting fake news, the text model is the basis, and the image information is used as an auxiliary. Meanwhile, image-text inconsistency is an issue that must be considered in bi-modal fusion. Therefore, in our proposed method, the image information will go through a gate structure for information selection before each fusion. Before the multi-modal feature fusion, the first inconsistency measure is conducted for filtering image features fed by the VGG-19. Then, we used co-attention to catch the relationship between text features and filtered image features. In the second fusion, we control the results after the first fusion through another inconsistency measure gate and then re-fuse with the text information. Finally, the prediction results were obtained through a BiLSTM network with a full connected layer. The output is the tag of this news (true or false). The overall structure of the model proposed is shown in Fig. 2. The model consists of: (1) The multi-modal feature extraction module, (2) The multi-modal feature fusion module, (3) The classification module.

3.1. The multi-modal feature extraction module

3.1.1. Text modality

By using traditional methods of word vector representation, for example, GloVe [41] or Word2Vec [42], the obtained word representation cannot change with the context, and cannot solve the problem of polysemy, but the pre-trained language model solves this problem to a certain extent. The BERT model [43] has achieved good results in 11 different natural language processing tasks and is considered to be milestone progress in the NLP (Natural Language Processing) field. In our model, we use BERT as the text encoder.

The initial input of the BERT model is a set of sentences $S = \{s_1, s_2, \ldots, s_m\}$, s_m represents the mth sentence, where $m \in M$; sentence s can be represented as a set of characters $s = \{w_1, w_2, \ldots, w_N\}$, w_n represents the nth character in the sentence, where $n \in N$. In our task, $S = x_i^T$, that is, the text part of the ith news. BERT input vector $E_n (n \in N)$ is composed of word embedding vector, segment embedding vector and position encoding vector. The word embedding vector is to find the corresponding vector representation of each word w_n according to the BERT embedding matrix. BERT can be trained in the form of sentence pairs, and the segmented embedding vector is used to identify the sentence. In the position encoding vector, BERT uses the learned position encoding to identify the position information of each word. Then, the BERT model uses Transformer [44] encoders to construct a multi-layer bidirectional network, which is stacked by multi-layer Transformer encoders. Each layer of the encoder is composed of a multi-head self-attention sub-layer and a feedforward neural network sub-layer.

For each news item, the process of extracting text features can be expressed by Eqs. (1)–(3).

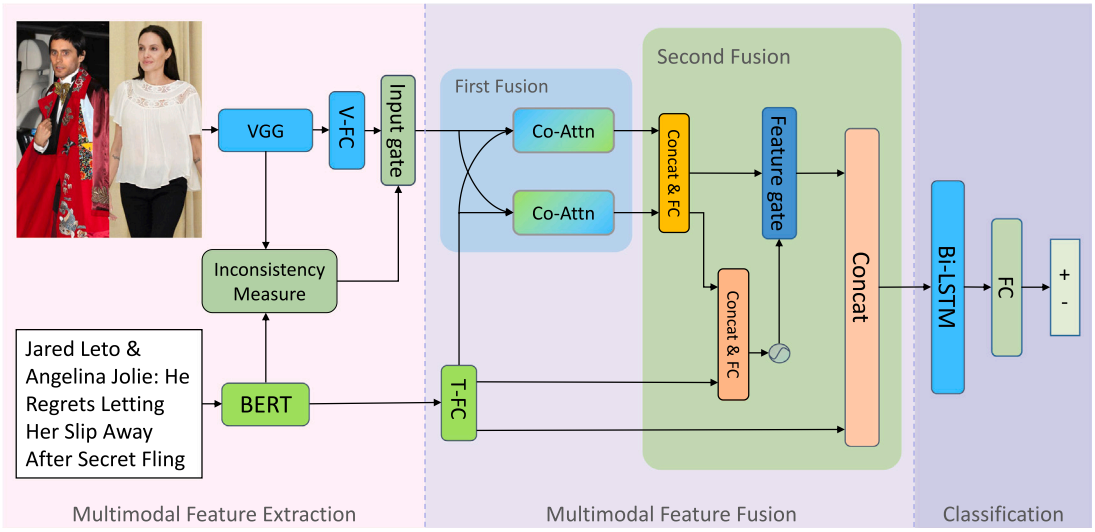


Fig. 2. Architecture of proposed model.

$$X_1^T = Token_{Bert}(S) \tag{1}$$

$$X_2^T = Sent_{Bert}(S) \tag{2}$$

$$m_t = \sigma(W_{t1} X_1^T) \tag{3}$$

Where S represents the text model x_i^T of ith news, $Token_{Bert}$ is the token-level output sequence of BERT model, and $Sent_{Bert}$ is the sentence-level output of BERT model, and $\sigma(\cdot)$ is the activation function. m_t are the output text features.

3.1.2. Image modality

VGG [45] is a typical CNN networks. VGG-19 is derived from the VGG architecture and consists of different layers, which are widely used in image analysis. VGG-19 has a total of 16 convolutional layers and 3 fully connected layers, in addition, there are 5 maximum pooling layers distributed under different convolutional layers. Due to the deepening of network structure, VGG-19 model has stronger learning ability in image feature extraction. In the VGG-19 network structure, the number of convolutional kernels starts from 64 in the first layer and gradually increases to 512, and then remains unchanged. In addition, due to the extensive use of small-sized convolutional kernels, the VGG-19 model usually requires fewer iterations to converge during training, thus improving the training speed.

For news x_i, the process of extracting image features can be expressed by Eqs. (4)–(6).

$$X_1^V = \widehat{VGG\text{-}19}(x_i^V) \tag{4}$$

$$m_v = \sigma(W_{v1} X_2) \tag{5}$$

$$X_2^V = \widehat{VGG\text{-}19}(\overline{VGG\text{-}19}(m_v)) \tag{6}$$

Here $\widehat{VGG\text{-}19}$ represents the feature extraction layers of VGG-19 model, $\overline{VGG\text{-}19}$ represents the average pooling layer of VGG-19 model, $\widehat{VGG\text{-}19}$ represents the classification layers of VGG-19 model. In order to align the dimensions in the calculation of the neural network layer, we also perform matrix shape transformation on the input of Eqs. (5)–(6).

3.1.3. The inconsistency measurement module

This module is used for inconsistency evaluation of text and image by measuring the semantic similarity between them. Unlike those works that measure consistency at the entity level [37,38], we think in terms of overall consistency. As shown in Fig. 1, the news image contains two person entities Jared Leto and Angelina Jolie, but it is obvious that they are formed by image stitching. If the consistency measurement is carried out from the entity level, it is easy to draw the conclusion that the graphics and texts are highly consistent, which is exactly the opposite of the actual situation. In the previous module, we used BERT and VGG-19 to learn the feature representation of text and image. Then we apply a sigmoid function on a linear layer, in which we concatenate the feature representation as the input, to measure the inconsistency between them. After that, we adopt a multiply gate to obtain the weighted image modality representation. Here, this module aims to measure the semantic consistency of image and text of the same news, and to further control the degree of information that can be fused through the following gate mechanism. Various methods can be used to measure semantic consistency. Through experimental comparisons, we found that a simple linear layer is a better choice for our dataset. For specific comparison experiments, see Section 4.7.

$$a = sigmoid(W_1[X_2^T; X_2^V]) \tag{7}$$

$$m_v = m_v * a \tag{8}$$

3.2. The multi-modal feature fusion module

Our multi-modal feature fusion is mainly done by two components: a Co-Attention [46] layer, which consists of two parallel Co-Attention blocks and a gate-based fusion module. Fig. 3 presents the structure of the Co-Attention block.

In the Co-Attention block, its query and key (=value) come from different places, i.e., if the query comes from the text, then the key (=value) comes from the image, and vice versa. For image modality, the calculation process is as follows.

$$h_e = \sigma([W_t m_t; W_i m_v]) \tag{9}$$

$$e = W_a h_e \tag{10}$$

$$A = \frac{exp(e)}{\sum exp(e)} \tag{11}$$

$$\overline{m_v} = \sum A W_v m_v \tag{12}$$

For textual modality, through the same calculation process, we can get text modal feature $\overline{m_t}$. By arranging Co-Attention block A and Co-Attention block B in parallel and combining them into a Co-Attention layer, Co-Attention block A uses text features as query and image

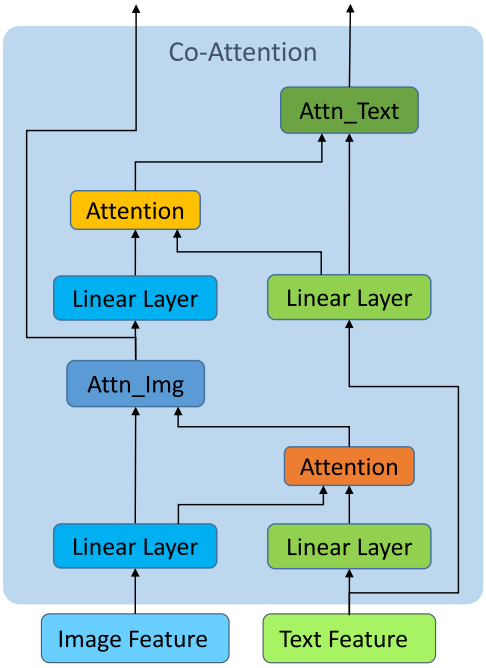


Fig. 3. Architecture of Co-Attention.

features as key; Co-Attention block B uses image features as query and text features as key. In this way, the interactive learning of information between text and image is achieved.

The gated-based fusion module is the second fusion. Different from the co-attention stacking component in existing work [47], our second fusion controls the feature information flow based on the gate mechanism again to obtain a more reliable feature representation. Wu et al. [47] use an iterative fusion of feature information, on the premise that the importance of bimodal information is equal. However, we believe that the image modality (and its fused features) is only auxiliary information, and its weight needs to be measured by the degree of consistency when it is gradually fused forward. See Section 4.7 for the contribution of this module. The fusion process is shown in Eqs. (13)–(16).

$$h_g = \sigma(W_{g1}[\overline{m_t}; \overline{m_v}]) \tag{13}$$

$$g = sigmoid(W_{g2}[m_t; h_g]) \tag{14}$$

$$m = g * h_g \tag{15}$$

$$c = [m_t; m] \tag{16}$$

3.3. The classification module

After the fused features are processed by the BiLSTM network and a feed forward layer, we can finally get the classification results, Binary cross entropy is adopted to define the loss function, as shown in Eq. (17).

$$L = -\sum [y \log \hat{y} + (1 - y) \log(1 - \hat{y})] \tag{17}$$

Where y is the true label of the news and $\hat{y}$ is the predicted label of the news.

4. Experiment

4.1. Datasets

To demonstrate the generalization of the proposed TRIMOON, we conduct experiments on two real-world datasets in different languages.

Table 1
The statistics of the datasets.

Dataset	Chinese	English
# of fake news	4748	3855
# of real news	4779	12844
# of images	9527	16699

The Chinese dataset is based on the Weibo dataset constructed by Jin et al. [48], which is a dataset widely used in this field. The authenticity of its news has been checked by an official rumour-refuting website.[1] The second one is the English dataset, which is the GossipCop sub-dataset of the FakeNewsNet [49] dataset. The real news in the English dataset are news articles scraped from E! Online,[2] which is a well-known website for publishing credible entertainment news. The fake news data comes from news stories with a score of less than 5 on the GossipCop website.[3] Table 1 shows the detailed statistics of the datasets. The training, developing and testing sets contain a number of news approximately with a ratio of 7:2:1 for both datasets.

4.2. Experiment settings

The parameters of the base model of text encoding and image encoding we use are frozen. The Bert-Base model is used for text encoding and the VGG-19 model is used for image encoding. When calculating the inconsistency score, we use the sentence-level vector output of Bert as the representation of the text sequence. Similarly, for the image modality, we use the features processed in the {0, 1, 3, 4}th layer of the VGG-19 classification module as the representation of the image. For the input in the first fusion, the output of the feature extraction layer of VGG-19 is used as the image modal feature, and the token-level output of Bert is used as the text representation for the text modality. The batch size is set to 16, the hidden layer dimension of text modality is 32, the maximum text input sequence length is limited to 512, the image region in VGG-19 is 49 (7*7), and the image feature dimension is 512. In order to reduce over-fitting, we follow each fully connected layer of the text modality with a Layer normalization layer, followed by a BatchNorm layer after each fully connected layer of the image modality, and then a dropout layer a rate of 0.6. Adam is used for optimization, with initial learning rate set to 1e-3, and training for a maximum of 100 epochs.

4.3. Baselines

To validate the performance of our proposed model, we compare it with two categories of baseline models: single-modality models and multi-modal models.

4.3.1. Single-modality models

Text: The model uses only text modal information in news for detection. We adopt the BERT pretrained model for word encoding in this model, then fed the word embedding sequence of a news into the BiLSTM to learn document-level feature representation, and then use a fully-connected layer with a softmax layer for classification prediction.

Visual: The visual model uses only images from news to classify them as fake or not. In this paper, the VGG-19 model pretrained on the ImageNet dataset is fine-tuned on our dataset for fake news detection.

[1] https://weibo.com/
[2] https://www.eonline.com/
[3] https://www.gossipcop.com/

4.3.2. Multi-modal models

VQA [50]: The task of Visual Question Answering (VQA) is to provide an accurate natural language answer given an image and an open-ended, natural language question about the image. We adopted the Visual QA model which was originally designed for a multi-class classification task to our binary classification task. This is done by replacing the final multi-class layer with a binary-class layer.

att-RNN [32]: This model is a recurrent neural network based on attention mechanism, which is used to fuse the features of three modalities of text, visual and social context. Among them, the text part is modelled by LSTM, and the image part is extracted by pre-trained VGG-19. For the fairness of the comparison, in the implementation, we remove the part dealing with social features

EANN [33]: The Event Adversarial Neural Network (EANN) is a kind of neural network based on event adversarial mechanism. By introducing an event classification as an auxiliary task, the model is guided to learn event-independent multi-modal features. The model uses TextCNN and pre-trained VGG-19 to extract text and visual modal features. Then it concatenate two modal features as the feature representation of fake news, which is fed into the fake news classifier and news events classifier.

MVAE [34]: This model is a multi-task model that combines a multi-modal variational autoencoder and a fake news detector. In this model, texts and images are extracted by BiLSTM and pre-trained VGG-19, respectively, and the concatenated features of the two are encoded as an intermediate representation for reconstructing input features and classifying fake news.

CARMN [15]: This model is mainly composed of two modules. One is called CARN (Cross-modal Attention Residual Network), and its role is to selectively extract information related to the target modality from the source modality while maintaining the unique information of the target modality. The other is called MCN (Multichannel Convolutional Neural Network), which is used to extract key information in text modalities from the CARN output.

4.4. Evaluation metrics

In this paper, we use Accuracy, Precision, Recall and F1 Score to measure model performance. Accuracy represents the percentage of correctly predicted results. Precision represents the ratio of the number of correctly named entities recognized by the model to the number of all identified named entities. Recall is the proportion of true positives to all correctly predicted outcomes. F1 score is the weighted harmonic mean of precision and recall. The formulae of the evaluation metrics are shown in Eqs. (18)–(21).

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP} \tag{18}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{19}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{20}$$

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{21}$$

where T and F denote the correctness of the prediction, which are correct and incorrect, respectively. P and N denote the prediction categories, which are positive and negative examples, respectively, and the result of summing these four values is the total number of samples.

4.5. Experimental results

To verify the validity our proposed model, we selected a set of uni-modal-based baseline models and a set of multi-modal-based baseline models for comparison. The comparative experimental results are shown in Table 2.

Table 2
Classification results on Chinese and English datasets.

	Method	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Chinese	text	**93.42**	72.70	81.77	80.55
	visual	87.87	70.29	78.10	77.00
	VQA	86.99	69.19	77.07	75.84
	att-rnn	89.91	84.36	87.05	87.51
	EANN	91.37	71.51	80.23	78.98
	MVAE	88.89	82.50	85.57	86.01
	CARMN	91.81	87.22	89.46	89.90
	TRIMOON	92.98	**88.83**	**90.86**	**91.26**
	Method	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
English	Text	**96.60**	84.67	90.24	83.82
	Visual	87.78	84.32	86.01	77.89
	VQA	87.31	84.72	86.00	77.89
	att-RNN	91.37	86.77	89.01	82.53
	EANN	87.74	86.17	86.95	79.60
	MVAE	91.91	86.09	88.91	82.23
	CARMN	94.24	87.50	90.74	85.11
	TRIMOON	96.25	**87.95**	**91.91**	**86.88**

Table 3
Ablation study on datasets.

		Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Chinese	Full Model	92.98	88.83	90.86	91.26
	w/o First Fusion	89.33	87.04	88.17	88.81
	w/o Second Fusion	90.79	86.49	88.59	89.08
English	Full Model	96.25	87.95	91.91	86.88
	w/o First Fusion	94.24	87.82	90.91	85.41
	w/o Second Fusion	94.70	87.71	91.07	85.62

On the Chinese dataset, TRIMOON outperforms other methods in Recall, F1, and Accuracy indicators, and is second only to the Text model in Precision indicator. Since our task is fake news detection, it focuses more on the Recall metric and F1 value. On both metrics, our model outperforms all other methods. Specifically, TRIMOON scored 92.98, 88.83, 90, 86, and 91.26 percentage on the Precision, Recall, F1 and Accuracy, respectively. For the two categories of models, the multi-modal models achieved generally higher performance than the single-modal models. TRIMOON scored 1.4 percentage points higher in F1 and 1.36 percentage points higher in Accuracy than the best-performing multi-modal model CARMN. Compared to the best-performing uni-modal model Text, our model has 12.48 percentage points higher F1 score and 10.71 percentage points higher Accuracy.

On the English dataset, similar conclusions can also be observed. Specifically, TRIMOON scored 1.17 percentage points higher in F1 than the best-performing multi-modal model, CARMN, and 1.77 percentage points higher in Accuracy. Compared to the best performing uni-modal model Text, TRIMOON achieved 1.67 percentage points higher F1 score and 3.06 percentage points higher Accuracy.

The experimental results show that our TRIMOON model has better performance than the existing models. Moreover, the multi-modal model outperforms the uni-modal model, which indicates that it is beneficial to incorporate image features into text features.

4.6. Ablation study

We conducted ablation experiments on both Chinese and English datasets to investigate the effectiveness of inconsistency detection in two fusion module. The complete model, the model after removing the first fusion and the model after removing the second fusion were used, respectively. The results of the experiments are shown in Table 3.

From the results, it can be seen that both fusion methods improve the performance of the model. Specifically, on the Chinese dataset, without the first fusion the performance in Precision, Recall, F1, Accuracy decreased by 3.65, 1.79, 2.69 and 2.45 percentage points respectively. And without the second fusion the performance in Precision,

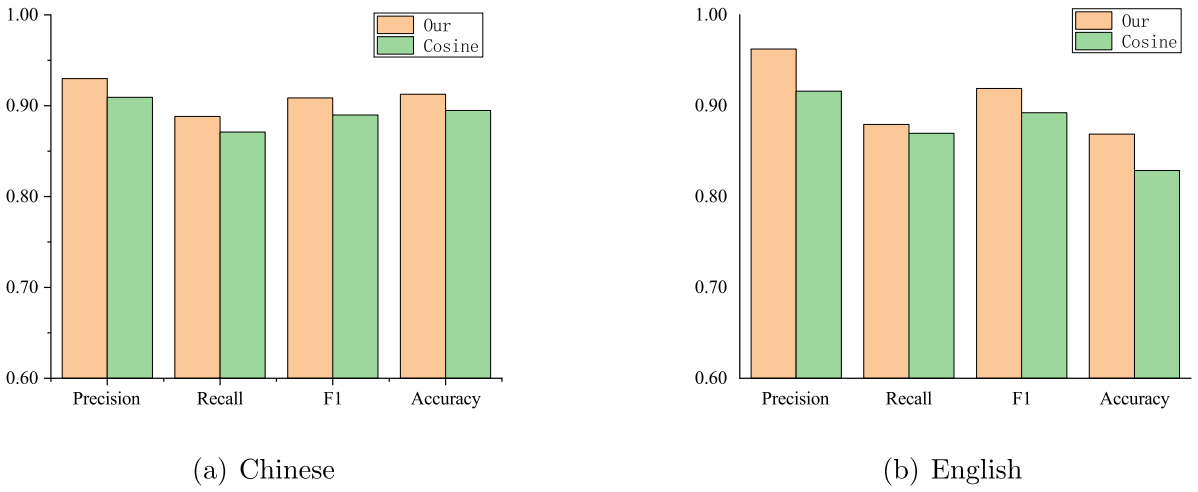


Fig. 4. Influence of Inconsistency module.

Recall, F1, Accuracy decreased by 2.19, 2.34, 2.27 and 2.18 percentage points respectively. On the English dataset, similar results can also be observed. The results indicate that our practice of performing two-round inconsistency-based multi-modal fusion to filter noise is effective, and the contribution of the first feature fusion is more obvious.

4.7. Influence of inconsistency module

A common calculation method for semantic consistency is cosine similarity. An implicit premise is that both parties involved in the computation are represented in the same semantic space. Since our task is the semantic coherence measure of text and picture modalities, there will inevitably be differences in semantic space between modalities. Therefore, we use a combination of linear layers and gating units to calculate semantic similarity. To verify the impact of the consistency measurement method on performance, we conduct a set of comparative experiments. Replace Eq. (7) with cosine similarity while the other components remain inactive. The specific experimental comparison results are shown in Fig. 4.

Our inconsistency module improves model performance on both datasets compared to using cosine similarity. Especially on English dataset, the improvement is more obvious, and the reason is that our method can better measure the inconsistency from the overall level of the event. This is also because of a major feature of the English dataset. Most of the image content in the dataset is about star events. The model must not only measure the consistency of characters but also measure the consistency of image and text from a global perspective rather than local semantic similarity.

4.8. Case study

In this section, we analyse the intuitive performance of our proposed approach on the dataset with real-life examples. Four typical fake news stories extracted from the test set are presented in Fig. 5. Our model is able to successfully detect the first three fake news stories, and the last one is a sample of wrong detection. The subject of text modal in Example 1 is that the sale of dog meat is illegal, while the images showing only sleeping dogs and babies do not illustrate dog meat and illegalities. Example 2 and Example 3 have the image content aligned with part of the text (or entities within it), but is unable to fully express the information consistent with the text modality. For instance, the avatar in Example 2 is Nicolas Cage himself, yet does not convey the semantics associated with his death as described by the text modality. Example 4 is an example of our model misdetection, where the image and text modalities are in good agreement at the coarse-grained level. This mistaken example also suggests a direction for future research into fine-grained image understanding and text consistency detection.

5. Discussion

By detecting the inconsistency between text modality and image modality, it allows our model to learn more reasonable feature representations when fusing multi-modal information, thus improving detection performance. In addition, the dominance of text modality in our fusion model is not overshadowed by image features, just as textual information is the dominant information representation for general news.

In fact, when images are inconsistent with the text, it is possible that the publisher is using inaccurate (misappropriated, historical) images as falsified evidence to support the conclusions in the news text. This is one of the reasons for constructing feature fusion representations based on inconsistent information. However, through the analysis of the above experimental results, we also found that our model is still at a coarse-grained level for image-text consistency detection, limited by the lack of semantic understanding of images. Therefore, finer-grained image understanding and image-text consistency detection is a challenge that still exists. In addition, when the image is highly consistent with the text, the features lose their diversity, in which case it is another challenging issue to exploit or enhance the features of the text modalities more efficiently.

6. Conclusion

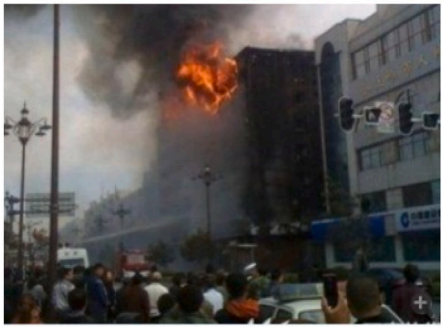
In this paper, a new multi-modal fusion method, a two-round multi-modal fusion method based on image-text inconsistency, is proposed to detect fake news. First, the vector representations of text modalities and image modalities are obtained through the BERT and VGG pre-trained models, respectively. Then we calculate the weight of the image modality for the detection task through the inconsistency measurement method and perform the first image-text fusion to obtain the fusion features. These features are not directly fed into the classifier, but is fused with the main features of the text modality again to obtain the final feature representation, and finally, the classifier is used for classification and detection. Compared with other methods, our approach has two obvious advantages: (1) before the initial fusion, the information is filtered through the inconsistency weight, so as to control the misleading of the content by the unreal images; (2) through the secondary fusion mechanism, to strengthen the dominance of text modal features while taking into account the image modal information to achieve better detection performance.

In the future research, further work can be carried out in the following areas. First, the design of image text representation and fine-grained consistency detection schemes based on vision-language

(a) It is illegal to sell dog meat.

(b) Hong Kong to start issuing $5,000 and $10,000 denomination notes next month.

(c) Nicholas Cage died in a skiing accident.



(d) The office of Northeast Securities in Gongzhuling was torched by angry shareholders.

Fig. 5. Some fake news samples. Example (a), (b) and (c) are correctly detected by our model, while Example (d) is incorrectly detected.

pre-trained models. Second, enhancing the learning capability of text features and inconsistency-based multi-modal fusion based on auxiliary approaches such as label representation, entity and event recognition. Third, considering semantic inconsistency measurement methods of different granularity. By using the above method to better integrate image and text content, obtain better fusion features, and achieve efficient and accurate detection of fake news.

CRediT authorship contribution statement

Shufeng Xiong: Methodology, Funding acquisition, Supervision. **Guipei Zhang:** Resources, Software. **Vishwash Batra:** Writing – review & editing. **Lei Xi:** Data curation, Formal analysis. **Lei Shi:** Writing – original draft, Validation. **Liangliang Liu:** Conceptualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by the MOE (Ministry of Education of China) Project of Humanities and Social Sciences (No. 19YJCZH198) the Science and Technology Planning Project of Henan Province, China (No. 222102110423) and the Natural Science Foundation of Henan Province, China (No. 222300420463). The author shall also be thankful to Pingdingshan University for providing computational resources for this work.

References

[1] A. Jain, A. Kasbe, Fake news detection, in: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS, 2018, pp. 1–5.

[2] X. Zhang, A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Inf. Process. Manag. 57 (2) (2020) 102025.

[3] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (5) (2020).

[4] N.T. Le, J.-W. Wang, D.H. Le, C.-C. Wang, T.N. Nguyen, Fingerprint enhancement based on tensor of wavelet subbands for classification, IEEE Access 8 (2020) 6602–6615.

[5] C. O'Connor, M. Murphy, Going viral: Doctors must tackle fake news in the covid-19 pandemic, BMJ (Online) 369 (2020) m1587.

[6] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on twitter during the 2016 u.s. presidential election, Science 363 (6425) (2019) 374–378.

[7] C. Matthews, How does one fake tweet cause a stock market crash, wall street & markets: Time, 2013.

[8] A.B. Prasetijo, R.R. Isnanto, D. Eridani, Y. Soetrisno, A. Sofwan, Hoax detection system on indonesian news sites based on text classification using SVM and SGD, in: 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE, 2017, pp. 45–49.

[9] M. Granik, V. Mesyura, Fake news detection using naive bayes classifier, in: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering, UKRCON, IEEE, 2017, pp. 900–903.

[10] L.M. Goyal, M. Mittal, J.K. Sethi, Fuzzy model generation using subtractive and fuzzy c-means clustering, CSI Trans. ICT 4 (2) (2016) 129–133.

[11] I.K. Sastrawan, I. Bayupati, D.M.S. Arsa, Detection of fake news using deep learning CNN–RNN based methods, ICT Express 8 (3) (2022) 396–408.

[12] M. Choudhary, S.S. Chouhan, E.S. Pilli, S.K. Vipparthi, Berconvonet: A deep learning framework for fake news classification, Appl. Soft Comput. 110 (2021) 107614.

[13] N. Rai, D. Kumar, N. Kaushik, C. Raj, A. Ali, Fake news classification using transformer based enhanced lstm and bert, Int. J. Cogn. Comput. Eng. 3 (2022) 98–105.

[14] R. Kumari, A. Ekbal, Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection, Expert Syst. Appl. 184 (2021) 115412.

[15] C. Song, N. Ning, Y. Zhang, B. Wu, A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks, Inf. Process. Manage. 58 (1) (2021) 102437.

[16] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, Detecting fake news by exploring the consistency of multimodal data, Inf. Process. Manage. 58 (5) (2021) 102610.

[17] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, H. Liu, Unsupervised fake news detection on social media: A generative approach, in: Proceedings of the AAAI Conference on49Artificial Intelligence, vol. 33, (no. 01) 2019, pp. 5644–5651.

[18] ZhouXinyi, ZafaraniReza, Network-based fake news detection, ACM SIGKDD Explor. Newsl. 21 (2) (2019) 48–60.

[19] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.

[20] P.H. Chen, C.J. Lin, B. Schölkopf, A tutorial on ν-support vector machines, Appl. Stoch. Models Bus. Ind. 21 (2) (2005) 111–136.

[21] N. Bhargava, G. Sharma, R. Bhargava, M. Mathuria, Decision tree analysis on j48 algorithm for data mining, Int. J. Adv. Res. Comput. Sci. Soft-Ware Eng. 3 (6) (2013).

[22] N. Ruchansky, S. Seo, Y. Liu, CSI: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017.

[23] J.C. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, E. Cambria, Supervised learning for fake news detection, IEEE Intell. Syst. 34 (2) (2019) 76–81.

[24] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of LREC 2022, 2022, pp. 3829–3839.

[25] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, K. Shu, Mining dual emotion for fake news detection, in: The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021, 2021, pp. 3465–3476.

[26] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.

[27] H. Guo, J. Cao, Y. Zhang, J. Guo, J. Li, Rumor detection with hierarchical social attention network, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 943–951.

[28] L. Wu, J. Li, X. Hu, H. Liu, Gleaning wisdom from the past: Early detection of emerging rumors in social media, in: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 99–107.

[29] M. Neuhaus, H. Bunke, A random walk kernel derived from graph edit distance, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition, SSPR, Springer, 2006, pp. 191–199.

[30] Q. Nan, J. Cao, Y. Zhu, Y. Wang, J. Li, MDFEND: Multi-domain fake news detection, in: International Conference on Information and Knowledge Management, Proceedings, 2021, pp. 3343–3347.

[31] G. Bathla, P. Singh, R.K. Singh, E. Cambria, R. Tiwari, Intelligent fake reviews detection based on aspect extraction and analysis using deep learning, Neural Comput. Appl. 34 (22) (2022) 20213–20229.

[32] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 795–816.

[33] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, 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, 2018, pp. 849–857.

[34] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

[35] H. Zhang, Q. Fang, S. Qian, C. Xu, Multi-modal knowledge-aware event memory network for social media rumor detection, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1942–1951.

[36] Y. Wang, S. Qian, J. Hu, Q. Fang, C. Xu, Fake news detection via knowledge-driven multimodal graph convolutional networks, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 540–547.

[37] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, Y. Yu, Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues, in: MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1212–1220.

[38] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, R. Ewerth, Multimodal analytics for real-world news using measures of cross-modal entity consistency, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 16–25.

[39] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, G. Xu, Entity-oriented multi-modal alignment and fusion network for fake news detection, IEEE Trans. Multimed. 24 (2022) 3455–3468.

[40] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, T. Lu, L. Shang, S. Li, Cross-modal ambiguity learning for multimodal fake news detection, in: Proceedings of the ACM Web Conference 2022, ACM, New York, NY, USA, 2022, pp. 2897–2905.

[41] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013) 3111–3119.

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

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., 2017, pp. 6000–6010.

[45] L. Geng, S. Zhang, J. Tong, Z. Xiao, Lung segmentation method with dilated convolution based on vgg-16 network, Comput. Assist. Surg. 24 (sup2) (2019) 27–33.

[46] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).

[47] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2560–2569.

[48] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 795–816.

[49] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (3) (2020) 171–188.

[50] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.