SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2023, Vol. 66 222102:1–222102:17
https://doi.org/10.1007/s11432-021-3530-7

# Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection

Yang YANG[1*], Ran BAO[2], Weili GUO[1], De-Chuan ZHAN[2], Yilong YIN[3] & Jian YANG[1]

[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094*, China;*
[2]*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China;*
[3]*School of Software, Shandong University, Shandong* 250101*, China*

Received 2 June 2021/Revised 1 January 2022/Accepted 31 May 2022/Published online 30 October 2023

**Abstract** With the development of the Internet, users can freely publish posts on various social media platforms, which offers great convenience for keeping abreast of the world. However, posts usually carry many rumors, which require plenty of manpower for monitoring. Owing to the success of modern machine learning techniques, especially deep learning models, we tried to detect rumors as a classification problem automatically. Early attempts have always focused on building classifiers relying on image or text information, i.e., single modality in posts. Thereafter, several multimodal detection approaches employ an early or late fusion operator for aggregating multiple source information. Nevertheless, they only take advantage of multimodal embeddings for fusion and ignore another important detection factor, i.e., the intermodal inconsistency between modalities. To solve this problem, we develop a novel deep visual-linguistic fusion network (DVLFN) considering cross-modal inconsistency, which detects rumors by comprehensively considering modal aggregation and contrast information. Specifically, the DVLFN first utilizes visual and textual deep encoders, i.e., Faster R-CNN and bidirectional encoder representations from transformers, to extract global and regional embeddings for image and text modalities. Then, it predicts posts' authenticity from two aspects: (1) intermodal inconsistency, which employs the Wasserstein distance to efficiently measure the similarity between regional embeddings of different modalities, and (2) modal aggregation, which experimentally employs the early fusion to aggregate two modal embeddings for prediction. Consequently, the DVLFN can compose the final prediction based on the modal fusion and inconsistency measure. Experiments are conducted on three real-world multimedia rumor detection datasets collected from Reddit, GoodNews, and Weibo. The results validate the superior performance of the proposed DVLFN.

**Keywords** multimodal learning, Wasserstein distance, rumor detection

**Citation** Yang Y, Bao R, Guo W L, et al. Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection. Sci China Inf Sci, 2023, 66(12): 222102, https://doi.org/10.1007/s11432-021-3530-7

## 1 Introduction

With the development of social media platforms, such as Twitter, Instagram, and Weibo, users can easily read and publish real-time posts. This advancement brings great convenience for governments and organizations to release real-time news and for users to keep abreast of surrounding events. For example, in the early stage of the COVID-19 pandemic, several reports on the global epidemic progress and related prevention suggestions were widely spread and rapidly forwarded on Weibo, Twitter, and other platforms. This greatly facilitated the control and prevention of the epidemic and had great social significance. However, various rumors emerged due to the convenience and openness of social media and caused a serious negative social impact [1, 2]. For example, rumors about presidential candidates emerged in the 2016 U.S. presidential election, which garnered significant attention in media and policy circles. Some journalists even claimed that the results of the 2016 election were a consequence of the rumors spread. In addition, rumors that amoxicillin can treat COVID-19 received a lot of attention and

* Corresponding author (email: yyang@njust.edu.cn)

 

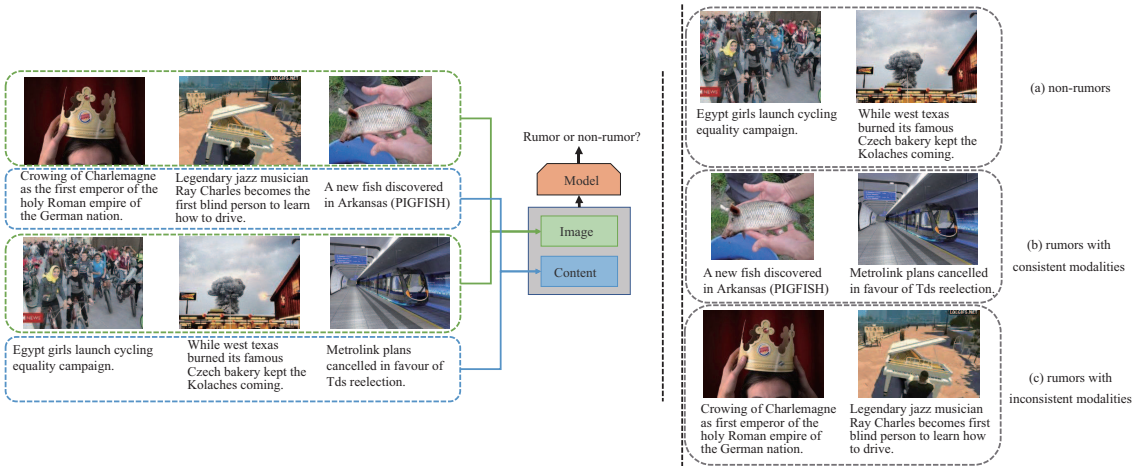


**Figure 1** (Color online) Multimodal detection framework (left) and intuition of the DVLFN (right). Multimodal detection framework always leverages possible visual-linguistic fusion for better performance. Meanwhile, rumors can be detected from two points: (a) non-rumor examples; (b) rumors with consistent modalities. The embeddings of the image and content are consistent, and we can verify the authenticity from visual or textual information. (c) Rumors with inconsistent modalities. Several rumors are difficult to verify the authenticity from any modality but are much easier to verify from the inconsistency degree of an image-text pair.

were widely reposted within a short time, which caused item shortage and public panic. Numerous fake posts on social media platforms may also greatly mislead consumers and even directly affect financial activities. These rumors have raised fears by misinforming users and corroding our society [3]. To solve this problem, fact-checking websites have been built to clarify rumors, such as snopes.com[1)] and politifact.com[2)]. However, they always rely on manual selection to detect rumors, which has obvious limitations on efficiency considering a large number of posts on social media. Therefore, the automatic detection of rumors is crucial to increasing the credibility of real-time posts.

As a matter of fact, posts on modern social media usually contain text and image information. With the rapid progression in computer vision [4,5] and natural language processing [6–8], we have confidence in detecting rumors from mass posts using artificial intelligence techniques. Initial approaches [9–12] usually concentrated on mining the text modality. They mainly employed handcrafted semantic features, such as term frequency-inverse document frequency, bag-of-words (BOW), or statistical features, to verify rumors, but these shallow features have a limited prediction ability. Inspired by the great success of deep neural networks, deep models have been recently exploited and demonstrated state-of-the-art performance. For instance, Ref. [13] employed recurrent neural networks (RNNs), i.e., long short-term memory (LSTM) and gated recurrent unit (GRU), to learn hidden features from text content for detecting rumors. Ref. [14] adopted text convolutional neural networks (CNNs) for obtaining local and global features from relevant posts. Meanwhile, image modality can depict an event. Hence, researchers have tried several deep visual approaches to analyze image modality, i.e., whether they are artificially generated or not. For instance, Ref. [15] extracted forensic features to evaluate the authority of the attached images. Ref. [16] fused multiple domain visual information for detection. Furthermore, inspired by the generative adversarial networks (GANs) [17], several approaches [18–20] modeled rumor detection as a text-based or image-based GAN-style framework to generate confusing training examples for enhancing the representation learning of rumor-indicative patterns. However, these methods focus on a single modality, i.e., image or text modality, but they ignore the fact that an effective multimodal fusion can obtain more discriminative embeddings and better predictions.

Based on this idea, researchers have attempted to combine two modal information in detection using multimodal fusion techniques. As shown in the left side of Figure 1, Refs. [21–24] deployed relative deep multimodal fusion networks for the rumor detection task. They usually directly fused multimodal deep embeddings or learned consistent embeddings and adopted a corresponding structural regularization or multitask loss for joint optimization. The basic assumption behind them is that either images or texts can verify rumors. These multimodal fusion methods mainly integrated two modalities' information, leaving the intermodal information without consideration. Figure 2 shows the inconsistency distribution

1) https://www.snopes.com/.
2) https://www.politifact.com/.

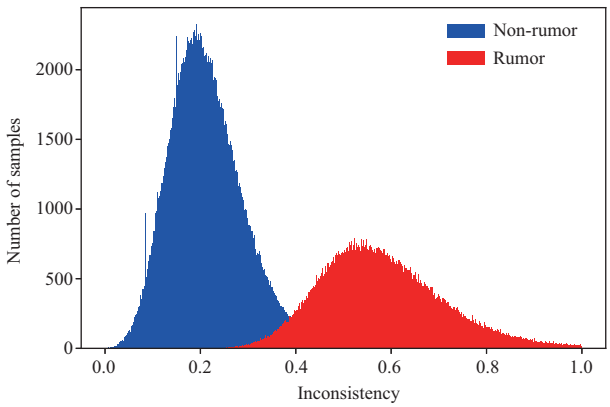


**Figure 2** (Color online) Inconsistency distribution of the Fakeddit dataset.

(calculated with Wasserstein matching) on Fakeddit dataset [25] examples. We find that most non-rumors have small inconsistency values, whereas rumors have large inconsistency values. Therefore, we can actually detect rumors from two aspects: (1) rumors with a consistent image-text pair, i.e., information coming from an image and text is consistent, and either an image or text can reflect the falsity. For example, as shown in Figure 1(b), the image and text depict the same event, whereas we can detect a rumor from the image/text or both. (2) Rumors with an inconsistent image-text pair, i.e., it is difficult to distinguish rumors from any modal information, but we can confirm rumors by detecting the mismatching of two modal contents. For example, as indicated in Figure 1(c), the text descriptions and attached images are mismatching. To sum up, existing multimodal methods aim to learn more discriminative embeddings by fusing all modal information, thus detecting rumors. Such an operator is applicable for the first type of rumors, but it cannot effectively detect the second type of rumors.

To this end, we propose a novel deep visual-linguistic fusion network (DVLFN), which includes intra-modal and intermodal evaluations, to comprehensively fuse multimodal information. Specifically, the DVLFN employs independently visual- and textual-based deep encoders to learn regional and global embeddings and then develops two evaluation criteria. (1) Modal aggregation. Through experimental verification, we use the early fusion (i.e., embedding fusion) to concatenate each modal embedding for prediction. (2) Intermodal inconsistency. The DVLFN calculates the Wasserstein distance between two modal region embeddings as an intermodal inconsistency. Consequently, the DVLFN combines the two measurements for the final prediction. Extensive experiments on three real-world datasets validate the superiority of using the DVLFN. Particularly, in terms of the F1 score, we achieve an absolute 2.7% improvement over the baselines on the NeuralNews dataset [26]. In summary, the contributions of this paper are as follows:

- We propose a novel DVLFN, which considers the extra intermodal inconsistency for rumor detection.
- We utilize visual and textual deep encoders to acquire global and regional embeddings and employ extra Wasserstein matching to comprehensively predict authenticity.
- In the experiments, our approach improves the performance on three real-world datasets, which validates that modal aggregation and intermodal knowledge are effective for enhancing the detector.

**Discussion.** Existing multimodal detection methods usually focus on designing effective fusion approaches for a good detection, and the method considering modal inconsistency only adopts manually designed inconsistent features or coarse-grained inconsistency. Actually, modal inconsistency is an effective criterion for rumor detection. To qualitatively detect modal inconsistency, we introduce the fine-grained inconsistent measure by Wasserstein matching, which aims to find subtler differences rather than simple aggregations. As a result, with modal fusion and modal inconsistency, we can acquire a good detection performance.

## 2 Related work

This paper aims to detect rumors by comprehensively considering the modal fusion and inconsistency information of posts. Therefore, our work is related to single-modal detection, multimodal detection, and cross-modal learning.

### 2.1 Single-modal detection

As a text is the main component of posts, traditional methods always utilize extracted semantic representations based on texts for rumor detection. For example, Refs. [9, 27] adopted BOW features from texts to extract relationships among posts and detect rumors. Ref. [12] employed a latent Dirichlet allocation (LDA) model to represent abstract semantic features for detection. However, these methods are limited to manually crafted features, which affect the detection performance. With the development of deep neural networks, several research studies have turned to utilize end-to-end text deep networks for the detection task. For example, Ref. [13] utilized the RNN model to learn high-level semantic embeddings of posts and predicted the possibility of authenticity. In addition to textual information, statistical features from texts, such as count of words, punctuation, hashtag topics(#), mentions(@), and URLs, were also utilized as auxiliary features [11, 28], which can capture the prominent statistical information for assisting detection. For example, Ref. [29] combined statistical features for detection, and Ref. [30] incorporated social contexts as auxiliary information into a hierarchical neural network via the attention mechanism for detection. Moreover, considering complementary visual information from attached images, several studies focused on distinguishing rumors by measuring the image quality. For example, Ref. [15] proposed to use visual forensic features for detection, and Ref. [16] designed a CNN-based network to automatically capture the complex patterns of forged images in the frequency domain. Recently, inspired by the GAN [17], several studies combined text- or image-based GANs to learn more robust embeddings. For example, Ref. [18] extracted co-occurrence matrices on three-color channels in the pixel domain and trained a robust model using the deep CNN framework. Ref. [19] proposed a text-based GAN-style approach, where the generator produced conflicting posts to pressurize the discriminator and learn strong rumor-indicative representations. However, all these methods are based on a single modality and cannot effectively fuse multimodal information contained in posts.

### 2.2 Multimodal detection

Recent studies aim to verify the credibility of multimedia posts by fusing multimodal information, i.e., texts and images. For example, Ref. [31] proposed the verifying multimedia task as a part of the MediaEval benchmark in 2015 and 2016. Multimodal deep neural networks are also proposed to fuse multimodal embeddings, in which CNNs or RNNs are always employed as a basic model for image/text modality. For example, Ref. [24] first incorporated deep neural networks for rumor detection by concatenating deep multimodal embeddings. Ref. [23] proposed an end-to-end event adversarial neural network to detect emerged rumors based on learning invariant multimodal embeddings. Ref. [22] trained a multimodal variational autoencoder jointly with a rumor detector to learn shared embeddings for texts and images. These methods always directly fuse multimodal embeddings or learn potentially consistent feature embeddings for detection, but they may receive interferences from noise information of indecipherable modality in inconsistent cases.

### 2.3 Cross-modal learning

Our work employs the inconsistency between two modalities, which is related to cross-modal learning. Currently, cross-modal learning aims to bridge connections among different modalities and has many applications, such as image captioning [32], cross-modal retrieval [33, 34], and visual question answering [35]. These tasks mainly utilize cross-modal consistency to construct the cross-modal generator or learn consistent embeddings. For example, state-of-the-art approaches in bidirectional image-sentence retrieval [36, 37] have leveraged visual-linguistic consistency to learn consistent embeddings and achieved great success on standard datasets, such as MSCOCO [38] and Flickr30K [39]. These methods usually adopted the modern distance or similarity measure to calculate the consistency degree between two modal global or regional embeddings. In this paper, we aim to utilize the cross-modal inconsistency in reverse. Refs. [26, 40, 41] mentioned the cross-modal inconsistency for detection, but these methods focused on manually designed inconsistent features or global inconsistency measure.

The remainder of this paper is organized as follows. Section 3 presents the proposed method, including the model, solution, and extension. Section 4 shows the experimental results on three rumor detection datasets under different settings. Section 5 concludes this paper.

# 3 Proposed method

We describe the details of our proposed DVLFN method in this section. The main goal of the DVLFN is to build a deep visual-linguistic fusion model for rumor detection, which not only integrates the intermodal inconsistency measure but also considers different modal fusions.

## 3.1 Notation

We first describe the notations used throughout this paper. Suppose there exist $N$ labeled posts with multimodal information, i.e., $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, $\boldsymbol{x}_i = \{\boldsymbol{x}_i^v, \boldsymbol{x}_i^w\}$ has two modalities, where superscript $v$ denotes image modality and $w$ denotes text modality. Note that $\boldsymbol{x}_i^w$ includes statistical information $\boldsymbol{x}_i^s$. $\boldsymbol{y}_i \in \{0,1\}^C$, where $C = 2$, $y_i = 1$ denotes a rumor; otherwise, it is a non-rumor.

## 3.2 Model pipeline

Without any loss of generality, the DVLFN can be divided into two important modules. (1) Embedding encoders, which learn discriminative embeddings for each modality. The key challenge is the selection of a multimodal encoder. In other words, we need to consider building an independent encoder for each modality or an interactive shared encoder for two modalities. In this study, we choose independent encoders for the following reasons. (a) Influence of embedding learning. The importance of different modalities is different, and thus a shared model may lead to a weak modality (i.e., modalities with weak classification performance), negatively influencing the embedding learning of the strong modality (i.e., modalities with strong classification performance) [42]. (b) Influence of the inconsistency measure. We focus on the inconsistency prediction, whereas the shared model actually tends to learn consistent embeddings, which may have a negative impact. Therefore, we develop visual and textual deep encoders, and we compare the two methods in the ablation study. (2) Classifier. We directly fuse global embeddings and then acquire the prediction with the fully connected (FC) network. Meanwhile, we adopt Wasserstein matching to measure the inconsistency calculation by region embeddings. The final prediction is constituted by the two results. The whole framework is shown in Figure 3. The image extracts raw global and regional features through the Faster R-CNN, whereas the text adopts the textual transformer encoder to learn global and regional embeddings. Ultimately, we aggregate the fused modal prediction and intermodal inconsistency for the final detection.

## 3.3 Basic networks

### 3.3.1 *Textual encoder*

In the upper-left corner of Figure 3, we first present the textual encoder $g_w$ for extracting text semantic embedding. Considering the success of BERT [43], we employ the transformer encoder for $g_w$. The input text is first tokenized into a token sequence according to WordPieces [44], followed by the standard BERT preprocessing method. We also add the special token [CLS] for learning global embeddings. Accordingly, the encoder can be represented as

$$\begin{aligned} \hat{\boldsymbol{e}}^w &= \mathrm{BERT}(\bar{\boldsymbol{e}}^w), \\ \bar{\boldsymbol{e}}^w &= \mathrm{LN}(\psi_{we}(\boldsymbol{x}^w) + \psi_{wl}(\boldsymbol{p}^w)), \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{e}}^w$ denotes output embeddings and $\bar{\boldsymbol{e}}^w$ represents the input of the BERT encoder. The representation for each sub-word token $\bar{\boldsymbol{e}}^w$ is obtained by summing up its word embedding and position embedding, followed by an LN (layer normalization). In detail, the instance can be represented as $\boldsymbol{x}^w = \{\boldsymbol{x}_l^w\}_{l=1}^{L_w}$, where $L_w$ denotes the instance length. The raw text representations $\boldsymbol{x}^w$ (initialized from input tokens) and position features $\boldsymbol{p}^w$ are fed through FC networks, i.e., $\psi_{we}$, $\psi_{wl}$, which project them into the same embedding space. Then, we sum the two features and use the LN to obtain $\bar{\boldsymbol{e}}^w$. We adopt BERT as the encoder to transform the embedding features layer by layer with adaptive attention weights.

The BERT encoder is a stack of identical blocks, which consists of multi-head attention and feed-forward layers, both wrapped in residual adds [8]. In detail, the $k$-th layer output embeddings can be represented as $\boldsymbol{z}^k = \{\boldsymbol{z}_l^k\}_{l=1}^{L_w}$, where $L_w$ denotes the input sequence length. We use sequence mask to

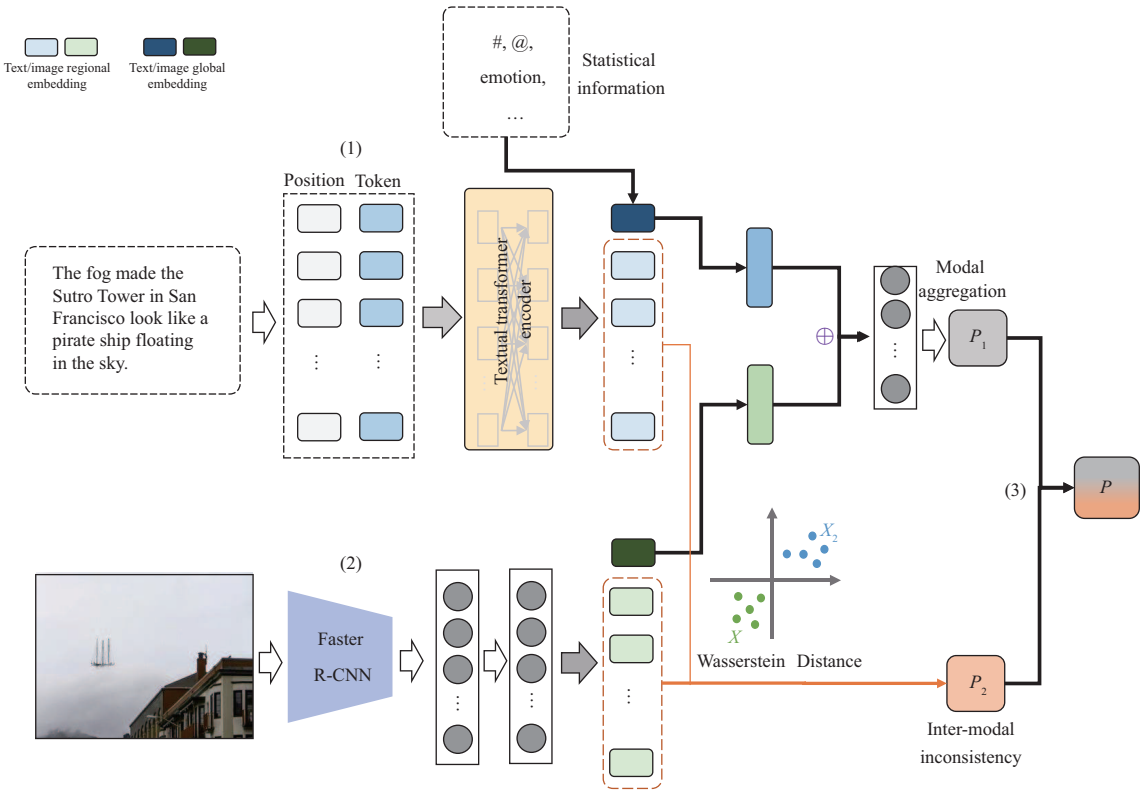


**Figure 3** (Color online) Illustration of the proposed DVLFN. It has three parts: (1) Textual model, i.e., textual transformer, fuses textual and statistical information to acquire semantic global and regional embeddings. (2) Visual model, i.e., Faster R-CNN, learns visual global and regional embeddings. (3) Prediction module includes intermodal inconsistency by calculating the Wasserstein distance and multimodal fusion through a generalized cross-entropy.

uniform the length of tokens to $L_w$. $\boldsymbol{z}^0$ is set as the input $\bar{\boldsymbol{e}}^w$. Then, the feature of the element $l$ in the $(k+1)$-th layer $\boldsymbol{z}_l^{k+1}$ can be computed by

$$\begin{aligned}
\bar{\boldsymbol{h}}_l^{k+1} &= \sum_{m=1}^{M} W_m^{k+1} \left( \sum_{j=1}^{L} A_{l,j}^m \cdot V_m^{k+1} \boldsymbol{z}_l^k \right), \\
\boldsymbol{h}_l^{k+1} &= \mathrm{LN}(\boldsymbol{z}_l^k + \bar{\boldsymbol{h}}_l^{k+1}), \\
\bar{\boldsymbol{z}}_l^{k+1} &= W_2^{k+1} \cdot \mathrm{GELU}(W_1^{k+1} \boldsymbol{h}_l^{k+1} + b_1^{k+1}) + b_2^{k+1}, \\
\boldsymbol{z}_l^{k+1} &= \mathrm{LN}(\boldsymbol{h}_l^{k+1} + \bar{\boldsymbol{z}}_l^{k+1}),
\end{aligned} \tag{2}$$

where $\boldsymbol{h}_l^{k+1}$ denotes hidden embeddings before the feed-ward operator in the $\{k+1\}$-th layer. Following the transformer encoder, we employ multi-head attention to jointly attend to information from different representation subspaces considering various positions, where $W_m^{k+1}$ denotes the weight of attention heads. The input features for each layer can be used to compute three matrices: $Q$, $K$, and $V$ corresponding to queries, keys, and values that drive the multi-head attention block, respectively. The dot-product similarity between queries and keys determines the attention distributions of values. In detail, $A_{l,j}^m$ denotes the attention weights between elements $l$ and $j$ in the $m$-th head, which is proportional to $(Q_m^{k+1} z_l^k)^{\mathrm{T}} (K_m^{k+1} z_j^k)$. $Q_m^{k+1}$, $K_m^{k+1}$, and $V_m^{k+1}$ are learnable weights for the $m$-th attention head. Then, weight-averaged values form the output of the attention block. $W_1^{k+1}$, $W_2^{k+1}$ and $b_1^{k+1}$, $b_2^{k+1}$ are learnable weights and biases in the feed-forward layer, respectively. We use the GELU activation instead of the ReLU operator following [45]. There is a residual structure after the two sub-blocks, followed by an LN layer.

Consequently, we can acquire the embedding of the input text from the [CLS] token and the word's embedding from other tokens. Moreover, the statistical features manually extracted from the text modality, i.e., $\boldsymbol{x}^s$ containing the hashtag topic, mention, and some semantic features, such as emotional polarity [24], are widely used in [24,28] as auxiliary information. Therefore, we concatenate the embedding of

the [CLS] token and the statistical features $\boldsymbol{x}^s$ as the new global text embedding, i.e., $\hat{\boldsymbol{e}}_0^w = [\hat{\boldsymbol{e}}_0^w, \boldsymbol{x}^s]$.

#### 3.3.2 *Visual encoder*

In the left-bottom corner of Figure 3, we present the visual encoder $g_v$ for extracting the image semantic embedding. The state-of-the-art Faster R-CNN [46] can obtain regional embedding and perform very well on the object detection task. Therefore, we adopt the Faster R-CNN as the visual encoder. Then, we utilize ResNet-101 to train the region proposal network (RPN), which is proven more powerful in feature extraction and performs better than VGG-16. In detail, we utilize the pre-trained Faster R-CNN to extract visual regions with pooled region of interest (ROI) embeddings for each image segment, denoted as $\{\hat{\boldsymbol{e}}_l^v\}_{l=1}^{L_v}$, where $l$ is the index, and $L_v$ is fixed for all image instances as [46] for a good performance. Therefore, the visual encoder can be formulated as

$$\hat{\boldsymbol{e}}^v = \text{Faster R-CNN}(\boldsymbol{x}^v), \tag{3}$$

where $\hat{\boldsymbol{e}}^v$ denotes the output embeddings, and $\boldsymbol{x}^v$ represents the input for Faster R-CNN. For image modality, we first resize the image to a fixed size for each instance before inputting it to the Faster R-CNN, i.e., $3 \times 224 \times 224$. For training, we first use ResNet-101 to generate a feature map, which is a stack of residual blocks based on deep convolutional layers. Then, we slide a small network over the feature map to obtain a series of low-dimensional features in the RPN. These features serve as a computing set of rectangular object proposals, each with an objectiveness score. We employ a well-trained RPN to obtain proposal ROIs for regional embedding. Similar to a textual encoder, we also encode the whole image for global embedding. Considering the excellent performance of ResNet-101 in training the RPN, we employ ResNet-101 to extract the global feature embedding.

### 3.4 Objective

Thus far, we have learned the text/image global and regional embedding: $\hat{\boldsymbol{e}}_0^w/\hat{\boldsymbol{e}}_0^v$ and $\{\hat{\boldsymbol{e}}_l^w\}_{l=1}^{L_w}/\{\hat{\boldsymbol{e}}_l^v\}_{l=1}^{L_v}$. $\hat{\boldsymbol{e}}_0^w$ and $\hat{\boldsymbol{e}}_0^v$ have different dimensions considering that $\hat{\boldsymbol{e}}_0^w$ combines the statistical information, whereas $\hat{\boldsymbol{e}}_l^w$ and $\hat{\boldsymbol{e}}_l$ have the same dimensions. We aim to use these output embeddings to make the final prediction from two aspects: (1) modal aggregation and (2) intermodal inconsistency.

Modal aggregation aims to effectively combine the two modal global information for prediction. In the experiment, we made three attempts: (a) embedding fusion, which adds or concatenates two modal embeddings and then inputs the fused embeddings to the FC network for prediction; (b) prediction fusion, which inputs each modal global embeddings to the independent FC network for prediction and then employs the max or mean operator to fuse two modal predictions; and (c) weighted fusion, which utilizes the attention mechanism on modal embeddings to calculate the weight for each modality and then combines the weight for the embedding fusion as [24] or prediction fusion as [47]. Based on the experimental verification, the embedding fusion gets the best results under the same settings on three datasets. A possible explanation is that the image contributes little to the detection, whereas the text information has more significance for detection. Therefore, the prediction or weight of the image modality may affect the final prediction more during the fusion process. Consequently, we employ embedding fusion in this paper. The modal aggregation can be formulated as

$$p_a = f(\psi_w(\hat{\boldsymbol{e}}_0^w) + \psi_v(\hat{\boldsymbol{e}}_0^v)), \tag{4}$$

where $\psi_w/\psi_v$ is the FC layer, which projects different modal global embeddings into the same embedding space, and $f(\cdot)$ denotes the classifier. Without any loss of generality, we utilize the FC networks here.

Intermodal inconsistency aims to measure the dissimilarity between two modalities. The direct ways include the following. (1) Global similarity (GS). We directly use the two modal global embeddings to calculate the similarity, i.e., $p_c = \cos(\psi_w(\hat{\boldsymbol{e}}_0^w), \psi_v(\hat{\boldsymbol{e}}_0^v))$. (2) Naive region similarity (NRS). We use a naive aggregation (e.g., average or sum) for all regional embeddings to compare GS, i.e., $p_c = \cos(\text{ave}(\mathcal{X}^w), \text{ave}(\mathcal{X}^v))$, where ave denotes the average operation. However, these kinds of methods mask important substructure differences. To overcome this problem, we turn to calculate the Wasserstein distance between the regional embedding of two modalities, which aims to find subtler differences rather than simple aggregations. We first illustrate the Wasserstein distance.

The Wasserstein distance [48, 49] is a function between probability distributions defined on a given metric space, which is more effective when the probability space has geometrical structures as compared

with Kullback-Leibler divergences, Hellinger distance, and total variation. Intuitively, the Wasserstein distance is the minimum cost of transporting the pile of one distribution into the pile of another distribution, which formulates the problem of learning the ground metric as minimizing the difference between two polyhedral convex functions over a convex set of distance matrices. Therefore, the Wasserstein distance is powerful in such situations by considering the pairwise cost.

**Definition 1** (Transport polytope)**.** For two probability vectors $r$ and $c$ in the simplex, $\Gamma(r,c)$ is the transport polytope of $r$ and $c$, namely, the polyhedral set of $p \times p$ matrices:

$$\Gamma(r,c) = \{\boldsymbol{T} \in \mathcal{R}_+^{p\times p} | \boldsymbol{T}\mathbf{1}_p = r, \boldsymbol{T}^{\mathrm{T}}\mathbf{1}_p = c\}.$$

**Definition 2** (Wasserstein distance)**.** Given a $p \times p$ cost matrix $\boldsymbol{M}$, the total cost of mapping from $r$ to $c$ using a transport matrix (or coupling probability) $\boldsymbol{T}$ can be quantified as $\langle \boldsymbol{T}, \boldsymbol{M}\rangle$. The Wasserstein distance problem is defined as

$$W(r,c) = \min_{\boldsymbol{T}\in\Gamma(r,c)} \langle \boldsymbol{T}, \boldsymbol{M}\rangle.$$

When $\boldsymbol{M}$ belongs to the cone of metric matrices $\mathbb{M}$, the value of $W(r,c)$ is the distance [50] between $r$ and $c$. In this case, assuming implicitly that $\boldsymbol{M}$ is fixed. That is, we can utilize the squared Euclidean distance of instances for calculating $\boldsymbol{M}$. $\boldsymbol{M}$ is a square matrix; otherwise, the corresponding position is filled with 0. Moreover, only $r$ and $c$ vary. We will refer to the optimal transport distance between $r$ and $c$. Notably, $W(r,c)$ is the cost of the optimal plan for transporting the predicted mass distribution $r$ to match the target distribution $c$. The penalty increases when more mass is transported over longer distances based on the ground metric $\boldsymbol{M}$. Therefore, with the region embeddings $\mathcal{X}^v$ and $\mathcal{X}^w$, $\boldsymbol{T}$ represents the node correspondences between $\mathcal{X}^v$ and $\mathcal{X}^w$. We define the matrix Wasserstein distance as follows.

**Definition 3** (Matrix Wasserstein distance)**.** Given two matrices $\mathcal{X} \in \mathcal{R}^{m\times d}, \mathcal{X}' \in \mathcal{R}^{n\times d}$, in which each row represents a region embedding. The matrix Wasserstein distance can be defined as $D_W(\mathcal{X}, \mathcal{X}') = W(\mathcal{X}, \mathcal{X}')$.

Here, different from traditional Wasserstein distance considering continuous probability distributions, we deal with finite sets of regional embeddings. Therefore, we can reformulate the Wasserstein distance as a sum rather than an integral and use the matrix notation commonly encountered in the optimal transport [51] to represent the transportation plan. Consequently, given two sets of vectors $\mathcal{X}^v = [\hat{\boldsymbol{e}}_1^v, \hat{\boldsymbol{e}}_2^v, \ldots, \hat{\boldsymbol{e}}_{L_v}^v]^{\mathrm{T}}$ and $\mathcal{X}^w = [\hat{\boldsymbol{e}}_1^w, \hat{\boldsymbol{e}}_2^w, \ldots, \hat{\boldsymbol{e}}_{L_w}^w]^{\mathrm{T}}$, we can define the Wasserstein distance as

$$\begin{aligned} &W(\mathcal{X}^v, \mathcal{X}^w) = \min_{\boldsymbol{T}\in\Gamma(\mathcal{X}^v,\mathcal{X}^w)} \langle \boldsymbol{T}, \boldsymbol{M}\rangle, \\ &\text{s.t. } \boldsymbol{T}\mathbf{1}_{L_v} = \mathbf{1}_{L_w}/L_w, \\ &\quad \boldsymbol{T}^{\mathrm{T}}\mathbf{1}_{L_w} = \mathbf{1}_{L_v}/L_v, \end{aligned} \tag{5}$$

where $\boldsymbol{M}$ is the distance matrix calculated by the two modalities of each instance, which contains the distances $s(\hat{\boldsymbol{e}}^v, \hat{\boldsymbol{e}}^w)$ between each element $\hat{\boldsymbol{e}}_l^v$ of $\mathcal{X}^v$ and $\hat{\boldsymbol{e}}_l^w$ of $\mathcal{X}^w$. We utilize the squared Euclidean distance for $s$ here. $\boldsymbol{T} \in \Gamma$ is a transport matrix (or joint probability), and $\langle\cdot,\cdot\rangle$ is the Frobenius dot-product. The total mass to be transported is equal to 1. Therefore, the row and column values of $\boldsymbol{T}$ must sum up to $\mathbf{1}_{L_w}/L_w$ and $\mathbf{1}_{L_v}/L_v$, respectively. The transport matrix $\boldsymbol{T}$ contains the fractions that indicate how to transport the values from $\mathcal{X}^v$ to $\mathcal{X}^w$ with the minimal total transport effort. In summary, we can define the inconsistency prediction as

$$p_c = 1 - \boldsymbol{e}^{-\gamma W(\mathcal{X}^v,\mathcal{X}^w)}, \tag{6}$$

where we employ the Laplacian kernel for calculating prediction and set $\gamma$ as 0.01 according to [52].

In summary, we can combine (4) and (6) for the final prediction

$$L = -\sum_{i=1}^{N} y_i \log(\max(p_{i,c}, p_{i,a})), \tag{7}$$

where the loss function can be represented as any convex loss function, and we adopt the cross-entropy here.

The parameters in (7) include $\boldsymbol{T}$ (transport matrix), $f$ (classifier of the modal fusion), and $g_v/g_w$ (image/text encoder). Actually, the solution of the transport matrix $\boldsymbol{T}$ has a closed form in the forward

process using $\boldsymbol{M}$. In detail, with the calculated $\boldsymbol{M}$, $\boldsymbol{T}$ is the solution of an entropy-smoothed optimal transport problem

$$\boldsymbol{T} = \arg\min_{\boldsymbol{T}} \langle \boldsymbol{T}, \boldsymbol{M} \rangle + \Omega(\boldsymbol{T}), \tag{8}$$

where $\Omega(\boldsymbol{T}) = -\frac{1}{\lambda}\sum_{mn} T_{mn} \log T_{mn}$ is the entropy of $\boldsymbol{T}$ and $\lambda > 0$ is the entropic regularization coefficient. $T_{mn}$ denotes the element in the $m$-th row and $n$-th column of $\boldsymbol{T}$. Based on the Sinkhorn theorem, we conclude that the transportation matrix can be written in the form of $T^{\star} = \mathrm{diag}(u)\boldsymbol{K}\mathrm{diag}(v)$, where $\boldsymbol{K} = \exp(-\lambda \boldsymbol{M})$ is the element-wise exponential of $\lambda \boldsymbol{M} - 1$. Moreover, in Sinkhorn iterations, $u$ and $v$ are updated continuously. Taking the $k$-th iteration as an example, the update is represented in the following form: $u = \frac{\mathbf{1}_{L_w}/L_w}{\boldsymbol{K}^{\mathrm{T}} u^{k-1}}$ and $v = \frac{\mathbf{1}_{L_v}/L_v}{\boldsymbol{K}^{\mathrm{T}} v^{k-1}}$. We adopt the gradient descent to update parameters considering the loss $L$. The details are shown in Algorithm 1.

**Algorithm 1** Code of the DVLFN

**Input**: Data: $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$; parameters: $\lambda$.
**Output**: Multimodal fusion network: $f, g_v/g_w$.

```
while stop condition is not triggered do
    for mini-batch sampled from D do
        Calculate p_a according to Eq. (4);
        Calculate the transport matrix T according to Eq. (8);
        Calculate p_c according to Eq. (6);
        Calculate L according to Eq. (7);
        Update model parameters using Adam;
    end for
end while
```

## 4 Experiments

In this section, extensive experiments are conducted based on three real-world datasets to demonstrate the effectiveness of our proposed method as compared with state-of-the-art methods.

### 4.1 Setups

#### 4.1.1 *Datasets*

We employ three commonly used real-world rumor detection datasets: Fakeddit [25], NeuralNews [26], and Weibo [24]. The Fakeddit dataset is a multimodal dataset crawled from Reddit of fake posts. It is collected from 22 different subreddits. The samples span over almost a decade and are posted on highly active and popular pages by over 300000 unique individual users, allowing to capture a wide variety of perspectives. The NeuralNews dataset is a constructed dataset, which consists of human and machine-generated articles with images and captions. The human-generated articles are sourced from the GoodNews dataset [53], which consists of New York Times news articles spanning from 2010 to 2018. Each news article contains a title, the main article body, and image-caption pairs. For fairness, we remove the image caption to keep the data format consistent with that in Fakeddit. The Weibo dataset is specifically constructed for rumor detection. In detail, considering the size and timeliness of the Weibo dataset in [24], we extend the original dataset by crawling more data from January 2018 to February 2019 with the same technique as [24]. Therefore, the Weibo dataset is an expanded dataset of the original paper. In this dataset, non-rumors are collected from authoritative news sources in China, such as Xinhua News Agency. The rumors are crawled from the website and verified by the official rumor debunking system of Weibo. This system encourages common users to report suspicious posts and examines suspicious posts by a committee of trusted users. We follow the same steps in [24]. We first remove the duplicated and low-quality images to ensure the quality of the entire dataset. Then, we apply a single-pass clustering method [27] to discover newly emerged events from posts. Finally, we split the whole dataset into training, validation, and testing sets in a 7:1:2 ratio and ensure that they do not contain any common event. The descriptions of the training and testing data distributions are listed in Table 1.

**Table 1** Statistics of three datasets

| Statistics | | Fakeddit | NeuralNews | Weibo |
|---|---|---|---|---|
| Training set | Rumor | 222081 | 22400 | 11792 |
| | Non-rumor | 341919 | 22400 | 11874 |
| Validation set | Rumor | 23320 | 3200 | 1678 |
| | Non-rumor | 36022 | 3200 | 1702 |
| Testing set | Rumor | 23507 | 6400 | 3371 |
| | Non-rumor | 35812 | 6400 | 3389 |
| All | | 682661 | 64000 | 33806 |

#### 4.1.2 *Implementation*

For the statistical features, we take the most commonly used social features in [24], which extracts 12-dimensional statistical features. The image encoder employs the Faster R-CNN [46], which uses fixed 36 ROIs. Each ROI and the global embedding are learned using ResNet-101 [54]. The text encoder employs the transformer [8], which has 12 layers of transformer blocks, where each block has 12 self-attention heads. The parameters are initialized from the BERT base, which is pre-trained on text data only. The model is trained with the following hyperparameters: ADAM optimizer [55] with a learning rate of 0.001, epoch number of 50, mini-batch size of 64, and weight decay of 0.001. The output layer size of each model is 200; i.e., each modal's final embedding is 200-dimensional. We implement the model using the MindSpore Lite tool[3)].

#### 4.1.3 *Comparison of methods*

To validate the proposed model on the rumor detection task, we compare our model with four groups of baseline methods, which are widely employed in the literature. (1) Single-modal methods: textual, visual, statistical, GAN-GRU [19], and MVNN [16]. (2) Direct multimodal early or late fusion methods: early fusion, late fusion, and weighted fusion. (3) Deep multimodal fusion methods: att-RNN [24], EANN [23], MVAE [22], VL-BERT [56], DRMM [57], MKN [21], SpotFake [58], and CARMN [59]. (4) Deep multimodal fusion method with inconsistency measure: DIDAN [26], SAFE [40], and EMFEND [41].

• **Textual.** A single deep model with texts only. We use BERT in this study.

• **Visual.** A single deep model with images only. We use ResNet101 in this study.

• **Statistical.** A single tree model with statistical features only. We use the SOTA decision tree model, i.e., LightGBM [60].

• **GAN-GRU.** A GAN-style approach. A generator is designed to produce augmented posts as rumors to pressurize the discriminator to learn strong rumor-indicative representations.

• **MVNN.** A deep approach that fuses the visual information of frequency and pixel domains for detecting fake news.

• **Early fusion.** Direct deep multimodal early fusion method. We employ visual and textual transformers as basic models and then add or concatenate different modal embeddings for final prediction, which is denoted as Add-EF/Concatenate-EF for simplicity.

• **Late fusion.** Direct deep multimodal late fusion method. We train each modal deep model independently and then utilize the mean/max operator for the final predictions, which is denoted as Mean-LF/MAX-LF for simplicity.

• **Weighted fusion.** Direct deep multimodal fusion method. We train the weight network to learn weights for the image and text and then utilize the weighted sum for modal embeddings or final predictions, which is denoted as Weighted-EF/Weighted-LF for simplicity.

• **VL-BERT.** A single flow-based multimodal transformer model, which concatenates the image-text input for joint training.

• **att-RNN.** A deep multimodal fusion network, which consists of an innovative RNN and an attention mechanism for fusing textual, visual, and statistical features.

• **EANN.** A deep multimodal fusion network, which employs a multitask loss, including rumor classification and event discrimination for learning common embeddings.

3) https://www.mindspore.cn/2020.

• **MVAE.** A deep multimodal fusion network, which trains a multimodal variational autoencoder jointly with a rumor detector.

• **DRMM.** A deep multimodal fusion network, which conducts deep interactions between images and sentences for modality feature aggregation.

• **DIDAN.** A deep multimodal fusion network considering inconsistencies, which manually designs a binary named entity indicator as an extra input.

• **MKN.** A deep multimodal knowledge-aware network considering multimodal content and external knowledge-level connections.

• **SAFE.** A deep multimodal fusion network considering cross-modal inconsistency, which computes multimodal relevance as the auxiliary objective for fake news detection.

• **SpotFake.** A deep multimodal fusion network that concatenates the textual and visual features obtained from pre-trained models.

• **CARMN.** A deep multimodal fusion network, which employs a cross-modal attention residual network for fusing multimodal features.

• **EM-FEND.** A deep multimodal fusion network considering cross-modal inconsistency, which models the complementary text information, mutual enhancement, and entity inconsistency.

To further investigate the impact of each item in the proposed model, we design several ablation studies for comparison. (1) w/o prediction: the variant of the DVLFN that removes the prediction term. (2) w/o inconsistency: the variant of the DVLFN that removes the inconsistency term. (3) w/o statistical: the variant of the DVLFN that removes the statistical information. (4) Different encoders: we also design various variants of the DVLFN with different visual/textual basic encoders, i.e., BERT+ResNet101, BERT+Vgg16, BERT+EfficientNet, LSTM+ResNet101, LSTM+Vgg16, LSTM+EfficientNet, GRU +ResNet101, GRU+Vgg16, and GRU+EfficientNet.

Considering that existing approaches have already mentioned the disadvantages of linear methods [22–24], we do not include linear methods in this study.

### 4.2 Classification results

In this subsection, we present the comparison results of the baselines and ablation methods with four criteria, i.e., accuracy, precision, recall, and F1 score. We only show the best results as [22, 24].

#### 4.2.1 *Comparison with the baseline methods*

Table 2 records the test results of all compared models. The results in the visual model are represented as "–" because there are no fake images in NeuralNews. NeuralNews is a constructed dataset that uses GLOVER [61] to generate fake articles with original titles and articles, and the raw images remain unchanged.

We draw the following observations. (1) The performance of texts is better than that of images on all datasets considering various criteria because texts always contain more semantic information for classification. Moreover, the distinct performance indirectly explains the inconsistency between images and texts. (2) The classification results of statistical information are also good, which validates the effectiveness of the statistical modality. (3) Among the different fusion methods, the performance of early fusion with the add operator performs the best on all the datasets considering various criteria. Therefore, we adopt it in the modal aggregation module. (4) The GAN-based model, i.e., GAN-GRU, improves slightly as compared with the original model, i.e., GRU. (5) The EANN and MVAE are not even as good as the single-modal models on partial settings because of the following reasons. (a) The EANN needs extra supervision information for multitask training, which is missing on both datasets. (b) The MVAE maps images and texts to the same vector space for a consistent constraint, but it brings confusion for embedding learning on the dataset (i.e., NeuralNews) with a large number of inconsistent instances. Hence, the MVAE performs worse on the NeuralNews dataset. (6) VL-BERT performs worse than independent training approaches (e.g., DVLFN, late fusion, and early fusion), which validates the advantage of the independent encoder for each modality. (7) The SOTA deep multimodal fusion method, i.e., DRMM, performs better than the single-modal and direct fusion models because they adopt a sophisticated integration strategy. (8) The fusion model considering inconsistency, i.e., DIDAN, performs well on the constructed dataset (i.e., NeuralNews), but it performs worse on real datasets, which indicates that the generalization of artificially constructed inconsistent features is not good. (9) The DVLFN performs better than the DIDAN and EM-FEND, which validates that the adaptively learned inconsistency is even better.

**Table 2** Classification results of different methods on three datasets[a)]

| Methods | Fakeddit | | | | NeuralNews | | | | Weibo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Pre | Rec | F1 | Acc | Pre | Rec | F1 | Acc | Pre | Rec | F1 |
| Textual | 0.868 | 0.895 | 0.886 | 0.890 | 0.930 | 0.920 | 0.942 | 0.931 | 0.959 | 0.956 | 0.963 | 0.959 |
| Visual | 0.794 | 0.842 | 0.811 | 0.826 | – | – | – | – | 0.847 | 0.851 | 0.837 | 0.844 |
| Statistical | 0.810 | 0.838 | 0.851 | 0.844 | 0.819 | 0.820 | 0.816 | 0.818 | 0.909 | 0.907 | 0.924 | 0.891 |
| Mean-LF | 0.891 | 0.907 | 0.914 | 0.910 | 0.930 | 0.933 | 0.927 | 0.930 | 0.955 | 0.954 | 0.955 | 0.955 |
| Max-LF | 0.892 | 0.908 | 0.912 | 0.910 | 0.931 | 0.932 | 0.930 | 0.931 | 0.945 | 0.952 | 0.936 | 0.944 |
| Weighted-LF | 0.890 | 0.903 | 0.917 | 0.910 | 0.924 | 0.922 | 0.928 | 0.925 | 0.947 | 0.961 | 0.932 | 0.946 |
| Add-EF | 0.901 | 0.920 | 0.916 | 0.918 | 0.933 | 0.940 | 0.926 | 0.933 | 0.962 | 0.969 | 0.954 | 0.961 |
| Concatenate-EF | 0.897 | 0.919 | 0.911 | 0.915 | 0.930 | 0.934 | 0.925 | 0.929 | 0.961 | 0.964 | 0.957 | 0.960 |
| Weighted-EF | 0.895 | 0.908 | 0.919 | 0.914 | 0.930 | 0.943 | 0.915 | 0.928 | 0.960 | 0.971 | 0.948 | 0.959 |
| GAN-GRU | 0.866 | 0.881 | 0.899 | 0.890 | 0.892 | 0.877 | 0.913 | 0.895 | 0.964 | 0.968 | 0.959 | 0.963 |
| MVNN | 0.768 | 0.828 | 0.778 | 0.802 | – | – | – | – | 0.842 | 0.849 | 0.826 | 0.838 |
| att-RNN | 0.901 | 0.916 | 0.921 | 0.918 | 0.893 | 0.881 | 0.909 | 0.895 | 0.964 | 0.972 | 0.955 | 0.964 |
| EANN | 0.871 | 0.898 | 0.887 | 0.892 | 0.857 | 0.872 | 0.836 | 0.854 | 0.953 | 0.966 | 0.939 | 0.952 |
| MAVE | 0.884 | 0.911 | 0.896 | 0.903 | 0.891 | 0.884 | 0.901 | 0.892 | 0.964 | 0.980 | 0.948 | 0.964 |
| VL-BERT | 0.888 | 0.909 | 0.907 | 0.908 | 0.921 | 0.941 | 0.898 | 0.919 | 0.946 | 0.957 | 0.934 | 0.945 |
| DRMM | 0.906 | 0.929 | 0.915 | 0.922 | 0.947 | 0.967 | 0.927 | 0.946 | 0.968 | 0.968 | 0.966 | 0.967 |
| MKN | 0.884 | 0.898 | 0.912 | 0.905 | 0.892 | 0.906 | 0.872 | 0.889 | 0.959 | 0.970 | 0.948 | 0.959 |
| SpotFake | 0.897 | 0.915 | 0.915 | 0.915 | 0.932 | 0.930 | 0.935 | 0.932 | 0.959 | 0.967 | 0.953 | 0.960 |
| CARMN | 0.888 | 0.913 | 0.900 | 0.906 | 0.891 | 0.893 | 0.887 | 0.890 | 0.966 | 0.967 | 0.965 | 0.966 |
| DIDAN | 0.893 | 0.919 | 0.903 | 0.911 | 0.956 | 0.952 | 0.960 | 0.956 | 0.967 | 0.964 | 0.969 | 0.967 |
| SAFE | 0.886 | 0.912 | 0.897 | 0.905 | 0.884 | 0.890 | 0.876 | 0.883 | 0.963 | 0.974 | 0.953 | 0.963 |
| EM-FEND | 0.905 | 0.925 | 0.917 | 0.921 | 0.951 | 0.967 | 0.934 | 0.950 | 0.967 | 0.970 | 0.966 | 0.968 |
| DVLFN | **0.915** | **0.931** | **0.928** | **0.929** | **0.982** | **0.989** | **0.975** | **0.982** | **0.978** | **0.981** | **0.974** | **0.977** |

a) The best results are highlighted in bold. "Acc", "Pre", and "Rec" denote the accuracy, precision, and recall.

**Table 3** T-test results of different methods on three datasets

| Methods | Fakeddit | | | | NeuralNews | | | | Weibo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| Add-EF | 0.001 | 0.013 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 | 0.004 | 0.000 |
| att-RNN | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.006 | 0.001 | 0.000 |
| EANN | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 |
| MVAE | 0.000 | 0.006 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.008 | 0.000 |
| DIDAN | 0.000 | 0.003 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.002 | 0.013 | 0.001 | 0.001 |
| DRMM | 0.000 | 0.017 | 0.012 | 0.000 | 0.000 | 0.003 | 0.000 | 0.000 | 0.000 | 0.009 | 0.002 | 0.001 |
| MKN | 0.000 | 0.001 | 0.006 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.003 | 0.000 |
| SAFE | 0.000 | 0.001 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| SpotFake | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.002 | 0.000 | 0.000 |
| CARMN | 0.000 | 0.001 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 |
| EM-FEND | 0.001 | 0.012 | 0.007 | 0.001 | 0.000 | 0.002 | 0.001 | 0.000 | 0.002 | 0.015 | 0.003 | 0.000 |

The EM-FEND performs better than the DIDAN on most settings because the EM-FEND considers more OCR information. (10) The DVLFN outperforms other baseline methods in all cases. We attribute the superiority to the intermodal inconsistency measurement. In addition, considering the difficulty of different datasets, the DVLFN promotes limited effects on the Fakeddit dataset, i.e., approximately 0.7% on F1 score, and it significantly improves on NeuralNews, i.e., approximately 2.7%. (11) To validate the significance and prove the superiority of the DVLFN, we perform statistical significance tests. Table 3 records the results of comparing state-of-the-art methods. We find that our proposed DVLFN indeed outperforms other algorithms significantly. For example, the p-values of the DVLFN are always lower than 0.05. (12) We conduct more experiments on the raw Weibo dataset [24]. Considering the F1 score, the best comparison method, i.e., EM-FEND, acquires 0.901, and the DVLFN gets 0.919, which also validates the effectiveness of the DVLFN on small datasets.

**Table 4** Ablation studies of the DVLFN and variants on three datasets[a)]

| Methods | Fakeddit | | | | NeuralNews | | | | Weibo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| w/o statistical | 0.910 | 0.923 | 0.926 | 0.926 | 0.980 | 0.987 | 0.970 | 0.980 | 0.972 | 0.968 | 0.973 | 0.972 |
| w/o prediction | 0.877 | 0.896 | 0.901 | 0.899 | 0.971 | 0.984 | 0.957 | 0.970 | 0.929 | 0.949 | 0.906 | 0.927 |
| w/o inconsistency | 0.901 | 0.920 | 0.916 | 0.918 | 0.933 | 0.940 | 0.926 | 0.933 | 0.962 | 0.969 | 0.954 | 0.961 |
| GS | 0.906 | 0.921 | 0.924 | 0.922 | 0.938 | 0.942 | 0.934 | 0.938 | 0.963 | 0.960 | 0.966 | 0.963 |
| NRS | 0.906 | 0.928 | 0.916 | 0.922 | 0.946 | 0.928 | 0.966 | 0.947 | 0.965 | 0.977 | 0.952 | 0.964 |
| Hungarian | 0.911 | 0.926 | 0.927 | 0.927 | 0.976 | 0.983 | 0.970 | 0.976 | 0.964 | 0.970 | 0.957 | 0.963 |
| DVLFN | **0.915** | **0.931** | **0.928** | **0.929** | **0.982** | **0.989** | **0.975** | **0.982** | **0.978** | **0.981** | **0.974** | **0.977** |

a) The best results are highlighted in bold.

**Table 5** Ablation study of different encoders on three datasets[a)]

| | | Fakeddit | | | | NeuralNews | | | | Weibo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Text | Image | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| BERT | – | 0.868 | 0.895 | 0.886 | 0.890 | 0.930 | 0.920 | 0.942 | 0.931 | 0.959 | 0.956 | 0.963 | 0.959 |
| LSTM | – | 0.865 | 0.887 | 0.890 | 0.889 | 0.883 | 0.907 | 0.852 | 0.879 | 0.956 | 0.950 | 0.961 | 0.956 |
| GRU | – | 0.865 | 0.885 | 0.892 | 0.888 | 0.891 | 0.876 | 0.911 | 0.893 | 0.957 | 0.961 | 0.953 | 0.957 |
| – | ResNet101 | 0.794 | 0.842 | 0.811 | 0.826 | – | – | – | – | 0.847 | 0.851 | 0.837 | 0.844 |
| – | Vgg16 | 0.766 | 0.809 | 0.801 | 0.805 | – | – | – | – | 0.840 | 0.882 | 0.779 | 0.827 |
| – | EfficientNet | 0.738 | 0.788 | 0.775 | 0.782 | – | – | – | – | 0.834 | 0.831 | 0.831 | 0.831 |
| BERT | ResNet101 | 0.906 | 0.921 | 0.924 | 0.922 | 0.938 | 0.942 | 0.934 | 0.938 | 0.963 | 0.960 | 0.966 | 0.963 |
| BERT | Vgg16 | 0.902 | 0.923 | 0.915 | 0.919 | 0.937 | 0.937 | 0.937 | 0.937 | 0.960 | 0.964 | 0.956 | 0.960 |
| BERT | EfficientNet | 0.899 | 0.914 | 0.920 | 0.917 | 0.932 | 0.938 | 0.925 | 0.931 | 0.955 | 0.959 | 0.949 | 0.954 |
| LSTM | ResNet101 | 0.903 | 0.919 | 0.921 | 0.920 | 0.890 | 0.891 | 0.888 | 0.889 | 0.957 | 0.961 | 0.953 | 0.957 |
| LSTM | Vgg16 | 0.895 | 0.916 | 0.909 | 0.913 | 0.894 | 0.901 | 0.885 | 0.893 | 0.954 | 0.956 | 0.952 | 0.954 |
| LSTM | EfficientNet | 0.896 | 0.910 | 0.918 | 0.914 | 0.890 | 0.899 | 0.879 | 0.889 | 0.953 | 0.965 | 0.939 | 0.952 |
| GRU | ResNet101 | 0.897 | 0.916 | 0.912 | 0.914 | 0.894 | 0.895 | 0.894 | 0.894 | 0.961 | 0.973 | 0.947 | 0.960 |
| GRU | Vgg16 | 0.890 | 0.916 | 0.900 | 0.908 | 0.896 | 0.901 | 0.890 | 0.895 | 0.960 | 0.963 | 0.956 | 0.959 |
| GRU | EfficientNet | 0.884 | 0.901 | 0.908 | 0.904 | 0.894 | 0.909 | 0.877 | 0.892 | 0.954 | 0.972 | 0.934 | 0.953 |
| DVLFN | | **0.915** | **0.931** | **0.928** | **0.929** | **0.982** | **0.989** | **0.975** | **0.982** | **0.978** | **0.981** | **0.974** | **0.977** |

a) The best results are highlighted in bold.

#### 4.2.2 *Ablation study*

**Study of variants of the DVLFN.** The results are shown in Table 4, from which we can conclude the following. (1) Using the inconsistency measure alone, i.e., w/o prediction, can improve the results. For example, the w/o prediction method achieves significant performance on the NeuralNews dataset, considering that NeuralNews is a constructed inconsistent dataset. (2) The w/o inconsistency method performs well on the three datasets, which validates the selection of the encoder. (3) The w/o statistical variant method is worse than the DVLFN but better than the w/o prediction and w/o inconsistency methods. This finding reveals that the statistical information is helpful to the prediction, but the effect is small.

**Study on different encoders.** To explore the effectiveness of different deep encoders for different modalities, we conduct more ablation studies. Related results are recorded in Table 5. For image and multimodal classifications, ResNet101 performed the best, followed by the other two encoders. Moreover, BERT achieved better results than the LSTM and GRU for text modality.

**Study on each loss term.** To explore the importance of the fusion term and inconsistency term, we adopt the weighted prediction replacing the max operation, i.e., $L = -\sum_{i=1}^{N} y_i \log((1-\beta)p_{i,a} + \beta p_{i,c}) + (1-y_i)\log(1-((1-\beta)p_{i,a} + \beta p_{i,c}))$. Figure 4 records the results of the weighted prediction (i.e., tuning the parameter $\beta$ in $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$) on the Fakeddit dataset. The results reveal the following. (1) Excessive $\beta$ will affect the performance. The reason is that content-based predictions are more important, and the inconsistency measure tends to detect rumors whose contents are difficult to classify. However, the modalities are obviously inconsistent. (2) The max operation achieves the best results because only when the modalities are extremely inconsistent will it be considered a rumor. Thus, the max operation ensures a better prediction.

**Study on different distance measures.** We conducted more studies to validate the role of intermodal

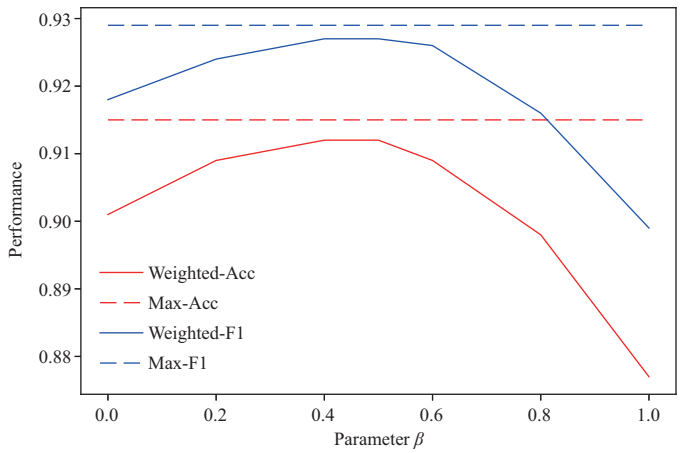


**Figure 4** (Color online) Performance of the DVLFN with different $\beta$ for weighted prediction.

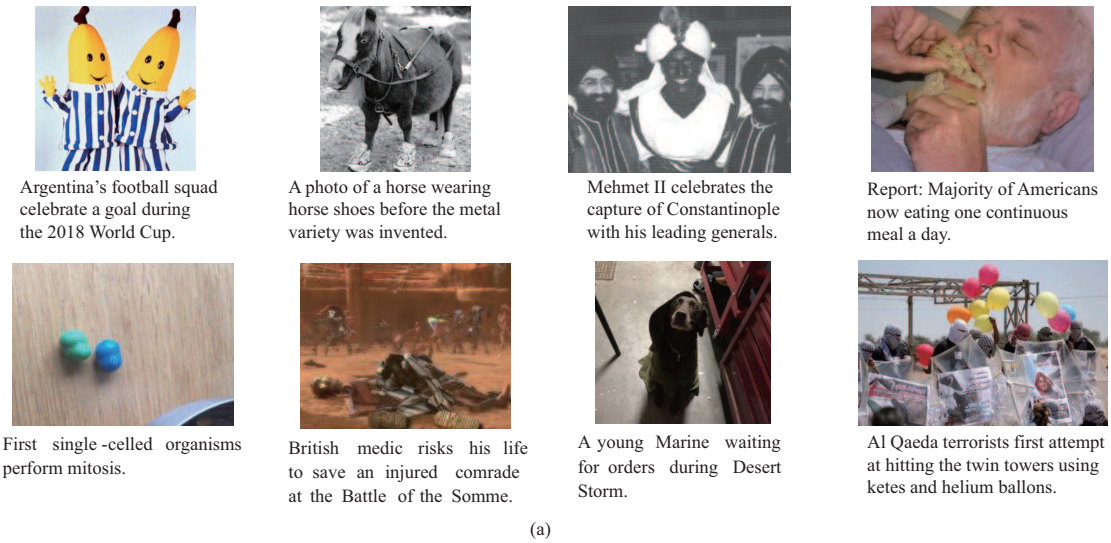


**Figure 5** (Color online) Several top rumors and non-rumors detected by the DVLFN from the Fakeddit dataset. (a) Examples of rumors; (b) examples of non-rumors.

inconsistency. (1) Different distance measures, i.e., GS and NRS. (2) Different graph matching methods, i.e., Wasserstein (i.e., DVLFN) and Hungarian. Table 4 records the results with different distance measures. The DVLFN achieves the best performance on various criteria of three datasets, which further

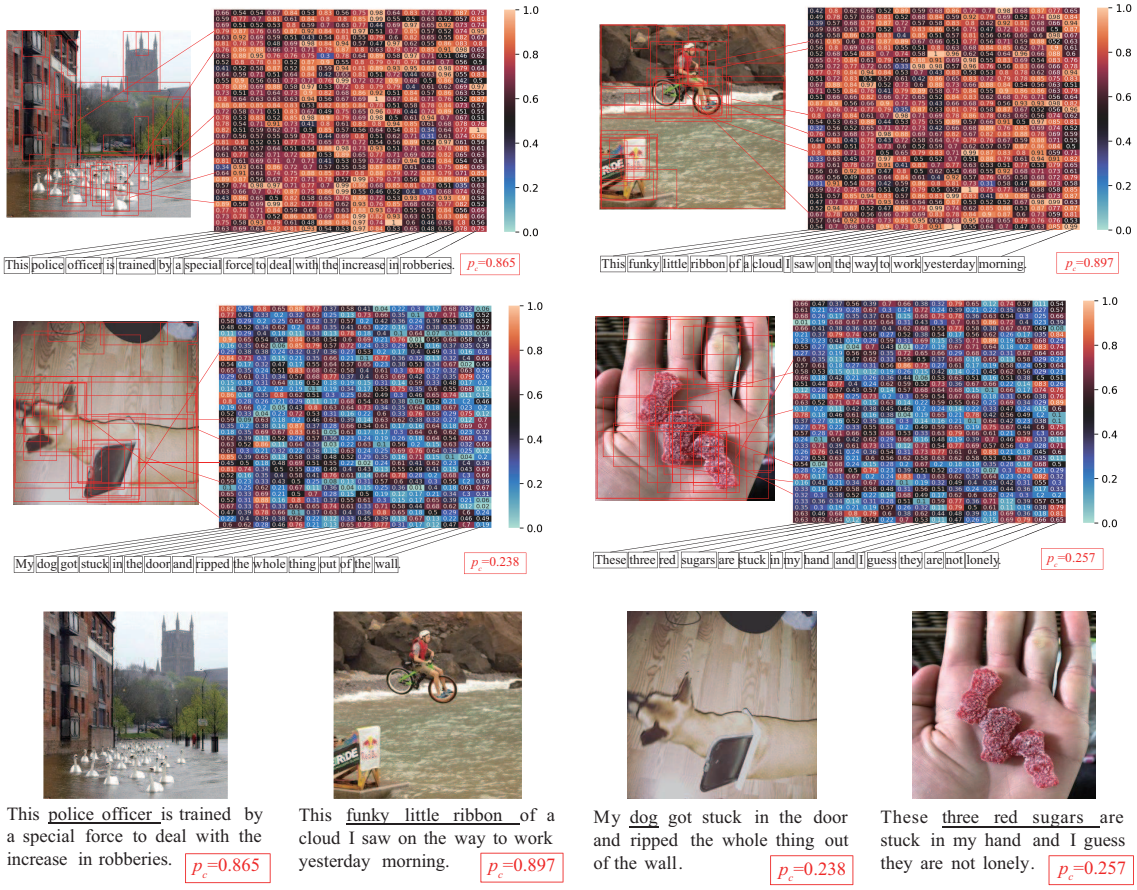


**Figure 6** (Color online) Examples with high inconsistency and low inconsistency scores detected by the DVLFN on the NeuralNews dataset. The two cases in the first row are multimodal examples with high inconsistency scores, and the two cases in the second row are multimodal examples with low inconsistency scores. The bottom row provides the raw image-text pairs for viewing convenience. The normalized matrix represents the distance matrix $M$ calculated by image regions and text words during the computation of inconsistencies. $p_c$ denotes the inconsistency score.

verifies the important role of considering region-based intermodal inconsistency.

### 4.3 Case study

We provide several successful examples for the qualitative analysis of the DVLFN. In detail, we sort the detected rumors and non-rumors based on the prediction scores of DVLFN and illustrate several demos from each class of the Fakeddit dataset in Figure 5. The portraits reveal that the DVLFN takes advantage of textual and visual content for rumor detection. For example, rumor examples carry significant fake patterns in text and image contents as compared with non-rumor ones. Furthermore, to illustrate the importance of inconsistency, we add cases to locate inconsistencies in images and texts according to the similarity matrix. Figure 6 indicates the results. The matrix represents the distance matrix $M$ calculated by the image regions and text words during the computation of inconsistencies. The values in the matrix are the normalized distances between image regions and text words, where large values denote dissimilarity. $p_c$ denotes the inconsistency score, where small values represent high consistency. The rumor cases with high inconsistency predictions are always with large regional distances. For example, in the first case in the top row of Figure 6, the image describes objects of "goose", "water", and "building", whereas the text describes objects of "police" and "robberies", The contents between the images and texts are inconsistent, and the distance matrix can effectively discover regions with large inconsistency values. On the contrary, consistent image-text pairs (i.e., the examples in the second row of Figure 6) can well locate the consistency of images and texts.

## 5 Conclusion

Rumors on social media have posed great challenges to supervisors. However, most of the current multimodal fusion methods ignored a notable characteristic of rumors, i.e., inconsistency between modalities. In this study, we develop the novel DVLFN, which detects rumors with the aid of a fine-grained inconsistency measure. In detail, the DVLFN first utilizes visual-linguistic deep embeddings to calculate regional embeddings, which aims to measure intermodal inconsistencies. Then, the DVLFN composes a reliable fusion mechanism to process concatenation for the final prediction. Experiments are conducted on real-world multimedia rumor detection datasets, and the results show the superior performance of the proposed DVLFN.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 62006118, 61906092, 61773198, 91746301), Natural Science Foundation of Jiangsu Province (Grant Nos. BK20200460, BK20190441), Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program, and CAAI-Huawei MindSpore Open Fund (Grant No. CAAIXSJLJJ-2021-014B).

**References**

1 Allport G W, Postman L. The Psychology of Rumor. New York: Russell&Russell Pub, 1947
2 Allcott H, Gentzkow M. Social media and fake news in the 2016 election. J Economic Perspect, 2017, 31: 211–236
3 Budak C. What happened? The spread of fake news publisher content during the 2016 U.S. presidential election. In: Proceedings of World Wide Web Conference, San Francisco, 2019. 139–150
4 Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 1915–1929
5 Yang Y, Zhan D C, Wu Y F, et al. Semi-supervised multi-modal clustering and classification with incomplete modalities. IEEE Trans Knowl Data Eng, 2021, 33: 682–695
6 Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference Machine Learning, Helsinki, 2008. 160–167
7 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, 2015
8 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 5998–6008
9 Gupt M, Zhao P, Han J. Evaluating event credibility on twitter. In: Proceedings of the SIAM International Conference on Data Mining, Anaheim, 2012. 153–164
10 Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media. In: Proceedings of the IEEE 13th International Conference on Data Mining, Dallas, 2013. 1103–1108
11 Wu K, Yang S, Zhu K Q. False rumors detection on sina weibo by propagation structures. In: Proceedings of the IEEE International Conference on Data Engineering, Seoul, 2015. 651–662
12 Jin Z, Cao J, Zhang Y, et al. News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, 2016. 2972–2978
13 Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the International Joint Conference on Artificial Intelligence, New York, 2016. 3818–3824
14 Yu F, Liu Q, Wu S, et al. A convolutional approach for misinformation identification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3901–3907
15 Boididou C, Papadopoulos S, Dang-Nguyen D T, et al. The certh-unitn participation@ verifying multimedia use 2015. In: Proceedings of MediaEval, 2015
16 Qi P, Cao J, Yang T, et al. Exploiting multi-domain visual information for fake news detection. In: Proceedings of the IEEE International Conference on Data Mining, Beijing, 2019. 518–527
17 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, Quebec, 2014. 2672–2680
18 Nataraj L, Mohammed T M, Manjunath B S, et al. Detecting GAN generated fake images using co-occurrence matrices. In: Proceedings of the Media Watermarking, Security, and Forensics, Burlingame, 2019
19 Ma J, Gao W, Wong K. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In: Proceedings of the World Wide Web Conference, San Francisco, 2019. 3049–3055
20 Jia B B, Zhang M L. Multi-dimensional classification via selective feature augmentation. Mach Intell Res, 2022, 19: 38–51
21 Zhang H, Fang Q, Qian S, et al. Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the ACM International Conference on Multimedia, Nice, 2019. 1942–1951
22 Khattar D, Goud J S, Gupta M, et al. MVAE: multimodal variational autoencoder for fake news detection. In: Proceedings of the World Wide Web Conference, San Francisco, 2019. 2915–2921
23 Wang Y, Ma F, Jin Z, et al. EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, 2018. 849–857
24 Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the ACM on Multimedia Conference, Mountain View, 2017. 795–816
25 Nakamura K, Levy S, Wang W Y. r/Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. 2019. ArXiv:1911.03854
26 Tan R, Plummer B A, Saenko K. Detecting cross-modal inconsistency to defend against neural fake news. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020. 2081–2106
27 Jin Z, Cao J, Jiang Y, et al. News credibility evaluation on microblog with a hierarchical propagation model. In: Proceedings of the IEEE International Conference on Data Mining, Shenzhen, 2014. 230–239
28 Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of the International Conference on World Wide Web, Hyderabad, 2011. 675–684

29 Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification. IEEE Trans Multimedia, 2016, 19: 598–608
30 Guo H, Cao J, Zhang Y, et al. Rumor detection with hierarchical social attention network. In: Proceedings of the ACM International Conference on Information and Knowledge Management, Torino, 2018. 943–951
31 Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at mediaeval 2015. In: Proceedings of the MediaEval 2015 Workshop, Wurzen, 2015
32 Karpathy A, Li F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3128–3137
33 Yang Y, Wu Y, Zhan D, et al. Deep robust unsupervised multi-modal network. In: Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, 2019. 5652–5659
34 Yang Y, Zhang C, Xu Y, et al. Rethinking label-wise cross-modal retrieval from a semantic sharing perspective. In: Proceedings of the International Joint Conference on Artificial Intelligence, 2021. 3300–3306
35 Wu Q, Teney D, Wang P, et al. Visual question answering: a survey of methods and datasets. Comput Vision Image Underst, 2017, 163: 21–40
36 Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 6077–6086
37 Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. ArXiv:2102.05918
38 Lin T, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context. In: Proceedings of the IEEE European Conference on Computer Vision, Zurich, 2014. 740–755
39 Huiskes M J, Lew M S. The MIR flickr retrieval evaluation. In: Proceedings of the ACM International Conference on Multimedia, British Columbia, 2008. 39–43
40 Zhou X, Wu J, Zafarani R. SAFE: similarity-aware multi-modal fake news detection. In: Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, 2020. 354–367
41 Qi P, Cao J, Li X, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In: Proceedings of ACM Multimedia, 2021. 1212–1220
42 Yang Y, Ye H, Zhan D, et al. Auxiliary information regularized machine for multiple modality feature learning. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 1033–1039
43 Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019. 4171–4186
44 Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. ArXiv:1609.08144
45 Hendrycks D, Gimpel K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. 2016. arXiv:1606.08415
46 Lee K, Chen X, Hua G, et al. Stacked cross attention for image-text matching. In: Proceedings of the European Conference Computer Vision, Munich, 2018. 212–228
47 Yang Y, Wang K, Zhan D, et al. Comprehensive semi-supervised multi-modal learning. In: Proceedings of the International Joint Conference on Artificial Intelligence, Macao, 2019. 4092–4098
48 Yossi R, Guibas L, Tomasi C. The earth mover's distance multi-dimensional scaling and color-based image retrieval. In: Proceedings of ARPA, 1997
49 Yang Y, Fu Z Y, Zhan D C, et al. Semi-Supervised multi-modal multi-instance multi-label deep network with optimal transport. IEEE Trans Knowl Data Eng, 2019, 33: 696–709
50 Villani C. Optimal Transport: Old and New. Berlin: Springer, 2008
51 Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval. Int J Comput Vision, 2000, 40: 99–121
52 Togninalli M, Ghisu M E, Llinares-López F, et al. Wasserstein weisfeiler-lehman graph kernels. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 6436–6446
53 Biten A F, Gómez L, Rusiñol M, et al. Good news, everyone! Context driven entity-aware captioning for news images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 12466–12475
54 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
55 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, San Diego, 2015
56 Su W, Zhu X, Cao Y, et al. VL-BERT: pre-training of generic visual-linguistic representations. In: Proceedings of the International Conference on Learning Representations, Addis Ababa, 2020
57 Tong M, Wang S, Cao Y, et al. Image enhanced event detection in news articles. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, 2020. 9040–9047
58 Singhal S, Shah R R, Chakraborty T, et al. SpotFake: a multi-modal framework for fake news detection. In: Proceedings of BigMM, Singapore, 2019. 39–47
59 Song C, Ning N, Zhang Y, et al. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manage, 2021, 58: 102437
60 Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 3146–3154
61 Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 9051–9062