Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Fake news detection for epidemic emergencies via deep correlations between text and images

Jiangfeng Zeng a, Yin Zhang b, Xiao Ma c,∗

a School of Information Management, Central China Normal University, Wuhan, China
b School of Information and Communication Engineering, University of Electronic Science and Technology of China, Shenzhen, China
c School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China

ARTICLE INFO

Keywords:
Epidemic diseases
Fake news detection
Semantic correlation
Multimodal fusion
Social networks

ABSTRACT

In recent years, major emergencies have occurred frequently all over the world. When a major global public health emergency like COVID-19 broke out, an increasing number of fake news in social media networks are exposed to the public. Automatically detecting the veracity of a news article ensures people receive truthful information, which is beneficial to the epidemic prevention and control. However, most of the existing fake news detection methods focus on inferring clues from text-only content, which ignores the semantic correlations across multimodalities. In this work, we propose a novel approach for Fake News Detection by comprehensively mining the Semantic Correlations between Text content and Images attached (FND-SCTI). First, we learn image representations via the pretrained VGG model, and use them to enhance the learning of text representation via hierarchical attention mechanism. Second, a multimodal variational autoencoder is exploited to learn a fused representation of textual and visual content. Third, the image-enhanced text representation and the multimodal fusion eigenvector are combined to train the fake news detector. Experimental results on two real-world fake news datasets, Twitter and Weibo, demonstrate that our model outperforms seven competitive approaches, and is able to capture the semantic correlations among multimodal contents.

1. Introduction

The breakout of major emergencies like epidemic diseases always breed a large quantity of disinformation, which greatly threatens the harmonious and sustainable development of the society (Al-Turjman & Deebak, 2020; Arafatur Rahman et al., 2020; Jiawei, Wang, & Liu, 2020; Kolhar, Al-Turjman, Alameen, & Abualhaj, 2020; Masoud & Mirmahaleh, 2020; Srivastava, Srivastava, Chaudhary, & Al-Turjman, 2020). As the proverb goes, when the truth was wearing shoes, the lie had spread all over the city. This well-known saying not only illustrates disinformation spreads quickly and widely like virus, but also reveals that disinformation causes a great harm to the human society. As reported in Hunt and Gentzkow (2017), an American Internet user was exposed to 1–3 pieces of fake news on average within the final month of the 2016 US presidential election. Foreign media has reported that online fake news will seriously influence the 2020 US presidential election. In the era of Internet, fake news widely spreading in social media networks may increase strains on the Internet via consuming a mass of Internet traffic. For instance, social media networks in different countries are flooded with kinds of fake news when COVID-19 (Corona Virus Disease 2019) rages all over the world, which wastes network resources so that people staying at home suffer a lot from the poor quality of Internet applications such as Netflix, YouTube, Tencent Meeting and so on. The widespread of fake news in recent years has verified the inability to defend against disinformation and the gap between AI driven counterfeiting techniques and detecting techniques is getting bigger and bigger in the Internet era (Lazer et al., 2018). Consequently, it is urgent to rebuild an information ecosystem that ensures authenticity and reduce strains on the Internet by curbing the propagation of fake news as soon as possible.

Theoretically speaking, fake news is committed to be worthless and rubbish. But in fact fake news, a mixture of truth and falsehood, is intentionally created to deceive or mislead readers and characterized by confusability. For example, it is observed that plenty of fake news about COVID-19 attracts more attention than the genuine news and to our surprise more people are willing to believe the fake ones. As a result, it poses tough challenges to check the veracity of a news article.

Automatically detecting fake news has been garnering an increasing

* Corresponding author.
E-mail addresses: jfzeng@ccnu.edu.cn (J. Zeng), yin.zhang.cn@ieee.org (Y. Zhang), cindyma@zuel.edu.cn (X. Ma).

https://doi.org/10.1016/j.scs.2020.102652

Available online 14 December 2020
2210-6707/© 2020 Elsevier Ltd. All rights reserved.
number of active research interest from the academic and industrial communities. Traditional approaches (Castillo, Mendoza, & Poblete, 2011; Feng, Banerjee, & Choi, 2012; Zhu et al., 2012) extract hand-crafted features from news textual content relying on expert knowledge, followed by traditional machine learning algorithms for training fake news classifier. These handcrafted features based methods are simple but lack comprehensiveness and flexibility (Kolhar, Al-Turjman, Alameen, & Abu-Alhaj, 2020; Wang et al., 2019). Substantial research have indicated that how to design artificial features is importantly critical to kinds of natural language processing tasks (Ma, Zeng, Peng, Fortino, & Zhang, 2019; Zhou, Zeng, Liu, & Zou, 2018). In the era of big data, highly distinguishable feature representation learning has become a new performance bottleneck of AI applications. The past decade has witnessed that deep neural networks including multi-layer nonlinear transformations are able to automatically learn more accurate and effective features from raw data and can be efficiently optimized via layer-wise gradient descent (Bahdanau, Cho, & Bengio, 2014; Jiang, Zeng, Zhou, Huang, & Yang, 2019; Shui-Hua & Zhang, 2020; Wang, Muhammad, Kong, Sangaih, & Zhang, 2020). Deep learning techniques bring in new insights into fake news detection and have achieved the state-of-the-art performance by automatically extracting features from news textual content (Chen, Li, Yin, & Zhang, 2018; Guo, Cao, Zhang, Guo, & Li, 2018a, 2018b; Jin, Cao, Zhang, & Luo, 2016; Li, Zhang, & Si, 2019; Lu & Li, 2020).

However, user-generated contents published on social media networks are always large-scale, noisy and multimodal. Multimodal machine learning gains an increasing number of attention, such as multimodal sentiment classification (Huang, Wei, Weng, & Li, 2020; Jiang, Wang, Liu, & Ling, 2020; Truong & Lauw, 2019), image captioning (Vinylas, Toshev, Bengio, & Erhan, 2015), visual question answering (Antol et al., 2015), multimedia analysis (Atrey, Hossain, El Saddik, & Kankanahalli, 2010) and so on. With the rapid development of social media networks, fake news evolves from classical text-only articles to articles attached with images or videos which carry richer information but pose more challenges to the detection task. It is apparent that multimodal news articles engage more readers than text-only articles. Some deep neural networks based methods have been investigated for multimodal fake news detection (Dhruv, Singh, Manish, & Vasudeva, 2019; Jin, Cao, Guo, Zhang, & Luo, 2017; Wang et al., 2018; Zhang, Fang, Qian, & Xu, 2019). Although these approaches mentioned above have proven to be successful and effective in detecting fake news, there still exist many difficulties in capturing the complex correlations between news textual content and news visual content.

In this work, we are dedicated to inferring newly emerged fake news by thoroughly analyzing multimodal news content that is fabricated and can be verified as false. In order to discover the complex correlations between textual modality and visual modality, we learn the multimodal fusion representation and the image-augmented text representation in a multi-task learning setting. The work most relevant to ours is NVAE (Dhruv et al., 2019), which learns a shared representation for multimodal news article by using VAE (Variational Autoencoder). However, there exist two main differences between our proposed method and NVAE. (1) In terms of one text article and several images attached to it, NVAE regards each text-image pair as one instance, which is unreasonable and unpractical. By contrast, our proposed method treats them as one instance. To be specific, VAE takes as input each text-image pair and outputs a unified representation of both textual modality and visual modality. Considering that simply treating all images as equally important and merging them equally to construct a fusion vector result in sub-optimal performance, we then leverage the attention mechanism to generate a multimodal fusion eigenvector. (2) To further mine the correlations between textual modality and visual modality, a hierarchical attention based approach is designed to learn image-augmented text representations, which is able to differentiate each word of the text document and select images of great importance to the prediction.

The key advantages of our proposed method are the following:

- In order to comprehensively mine the semantic correlations between textual modality and visual modality, we put forward a novel end-to-end deep neural networks based method, dubbed FND-SCTI which learns a multimodal fusion representation and an image-augmented text representation in a multi-task learning setting.
- Extracting multimodal correlations is core to checking whether a piece of multimodal news is fake. In one hand, we exploit images as supervisors to highlight the important part of news text. In the other hand, a shared multimodal representation is obtained by using a variational autoencoder. The fake news detector then take as input the multimodal fusion representation and the image-reinforced text representation to predict the news as fake or not.
- Compared with some competitive methods, qualitative experiments on two real-world datasets are conducted to validate the efficacy of our proposed approach for multimodal fake news detection and experimental results show the superiority of our proposed method.
- We demonstrate that our proposed FND-SCTI is afford to reveal the correlations between textual modality and visual modality by visualizing the attention weights of text-image pair.

The remainder of this paper is organized as follows. In Section 2, the related work of fake news detection is briefly introduced. Afterwards, we detailly illustrate the proposed FND-SCTI in Section 3. In Section 4, we conduct extensive experiments to demonstrate the superiority of our proposed method. Finally, we draw a conclusion and envision the future work in Section 5.

2. The related work

The emergence of social media networks not only makes it convenient for users to access, create and share messages, but also breeds and spreads misinformation like rumors and fake news which have a more negative impact on the human society than ever before. Over the past few decades, major emergencies have occurred frequently and result in the flooding of disinformation such as fake news. Fake news detection has obtained more and more attention recently. In this section, we provide a brief background on fake news detection and explore feasible insights to tackle fake news detection by modeling the multimodal news contents.

2.1. Unimodal fake news detection

The majority of existing approaches focus on detecting fake news by analyzing the text-only contents. Earlier methods manually extracted text features and use traditional machine learning algorithms as fake news detector. As a representative, Castillo et al. (2011) devised hand-crafted features like linguistic features, topic features and propagation features which were fed into a SVM based classifier for final prediction. Wu, Yang, and Zhu (2015) argued that ignoring the propagation structure of the posts resulted in a sub-optimal detection performance and attempted to capture the high-order propagation patterns which were first combined with content semantic features and then fed into a graph-kernel based SVM classifier. However, manually extracting features from text is complicated, time-consuming and costly, making it difficult to take advantage of big data.

Deep learning techniques have been fully investigated for fake news detection due to their ability of powerful representation learning. Ma et al. (2016) were first to devise a RNN-based method for rumor detection on microblogging platforms. In the work of Yu, Liu, Wu, Yang, and Tan (2017), Yu et al. attempted to identify misinformation by using a CNN-based method. Recently, Shu, Wang, and Liu (2019) explored the role of social context for textual fake news detection and put forward a tri-relationship embedding framework TriFN to represent the relationship among publishers, users and news articles, which proved the effectiveness of social context in improving the detection performance. Jin et al. (2016) tried to identify rumors by discovering conflicting user
comments. Guo, Cao, Zhang, Shu, and Liu (2019) pointed out that not only did the sentiment features extracted from user comments contribute to recognize misinformation but also the sentiment features extracted from news articles played an important role in detecting fake news. Ma, Gao, and Wong (2018a) designed two tree-structured recursive neural networks from bottom-up and top-down respectively and extracted content semantic features by following the propagation trees. Shu, Cui, Wang, Lee, and Liu (2019) used a co-attention mechanism to find top K important sentences of a news article and top K important user reviews as clues for final classification. Ma, Gao, and Wong (2018b) and Li et al. (2019) detected rumors and user stance jointly in a multi-task learning setting. Similar work like Cui, Wang, and Lee (2019), Liu and Wu (2018) and Zellers et al. (2019) have been proposed. Wang, Yang, et al. (2020) proposed a reinforced weakly supervised method leveraging users’ reports as weak supervision to enlarge the amount of training data. Lu (2020) developed GCAN which predicted whether a piece of news is fake or not given the short-text tweet and the corresponding sequence of retweet users without comments. In our work, we aims to address the early detection of disinformation by analyzing only the news contents since the social context is generated when the news gets widely propagated.

The development of image editing softwares like Photoshop and the big success achieved by Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Zeng, Ma, & Zhou, 2019) in image synthesis are reducing the technical threshold of image forging. Fake image detection has also been investigated recently. In the work of Zhou, Han, Morariu, and Davis (2018), Zhou et al. proposed a Faster R-CNN based method to extract tampering features from an RGB stream and a noise stream for image manipulation detection. Qi, Cao, Yang, Guo, and Li (2019) designed a framework including a frequency domain module, a pixel domain module and a fusion module to learn visual representations for fake image detection.

2.2. Multimodal fake news detection

With the rapid development of mobile intelligent devices and information technology, there are growing an increasing number of posts which are attached with images in social media networks. Unimodal information is insufficient and easy to be affected by various external factors. What’s more, due to the complementarity among different modalities, multimodal feature fusion is getting more and more attention (Baltrušaitis, Ahuja, & Morency, 2018). It is highly believed that fusing textual features and visual features plays a very important role in better understanding the semantics of raw data.

Recently more and more researchers started to deal with the problem of multimodal fake news detection. Jin et al. (2017) proposed Att-RNN exploiting a neural attention on top of the LSTM to fuse text features, social context and image features. Wang et al. (2018) devised a generative adversarial network (GAN) based model whose core is using the discriminator to remove any event-specific features. They exploited the event-invariant features learned to train the fake news detector. In the work of Dhruv et al. (2019), Khattar et al. leveraged a variational autoencoder (VAE) to learn a shared multimodal representation and achieved the current state-of-the-art results for multimodal fake news detection. The biggest difference between our work and MVAE is that our work not only learns the multimodal fusion eigenvectors but also enhances the representation for text content using visual aspect attention.

Though these approaches mentioned above have proven to be successful and effective in detecting fake news, the problem of predicting
fake news is far from being essentially solved. Especially, multimodal fake news detection methods still leave far space for improvement.

3. The proposed approach

As shown in Fig. 1, a high-level illustration of the proposed approach for multimodal fake news detection is clearly introduced. Specifically, we first formulate the problem of multimodal fake news detection in Section 3.1. Then an overview of our proposed model is presented in Section 3.2. Afterward, 4 components of our proposed model are detailledly illustrated in Sections 3.3–3.6, respectively. Finally, we present the training objectives to jointly train the fake news detector and multimodal variational autoencoder in Section 3.7. In order to make readers clear about the notations in this section, Table 1 is added to explain some important symbols.

3.1. Problem statement

We are given a set of news articles with manually annotated labels $D = \{x_1, y_1\}_{i=1}^n$ where $X = \{x_1\}_{i=1}^n$ denotes the news pieces and $Y = \{y_1\}_{i=1}^n \subset \{0, 1\}$ denotes the corresponding labels of 0 for true and 1 for fake. Suppose a piece of news article includes one text document $d$ and images attached with the post $img = \{img_1, img_2, ..., img_m\}$ that has $m$ images. The text document $d$ contains a sequence of $L$ sentences each of which has $l$ words, i.e., $s_i = \{w_1, w_2, ..., w_l\}, i \in [1, L]$. The task of multimodal fake news detection in our work aims to predict whether a piece of news article is fake or not depending on both the text document $d$ and images attached $img$. For example, in the first module of Fig. 1, we are given one piece of news which contains one text document and three images attached. We aim to propose one machine learning based model to predict whether this piece of news is fake or not. Since only news contents of textual modality and visual modality are utilized to prefer the veracity, our work is capable of tackling the early detection of fake news without the social context such as user comments, user profiles, propagation networks and so on.

3.2. Overview

The basic idea behind our work is that thoroughly mining the deep correlations between textual modality and visual modality plays a key role in understanding the content semantics and discovering the clues for preferring whether a piece of news article is fake or not. Fig. 1 presents the detailed architecture of the proposed FND-SCIT. It can be observed that FND-SCIT is composed of 4 modules, i.e., word embeddings and VGG features, image-enhanced text representation learning, multimodal representation learning and fake news detector. The first module digitizes the raw data, and learns word embeddings and VGG features for textual data and visual data respectively. Both of the second module and the third module are designed to discover the correlations between textual modality and visual modality from two different perspectives. Fake news detector takes as input the outputs of the second module and the third module, and is optimized in a multi-task learning setting. We elaborate the four modules in the following.

3.3. Word embeddings and VGG features

This module aims to prepare the inputs for the second module and the third module. It is known to all that each NLP application starts with transforming word tokens into word embeddings and the quality of word embeddings has a great influence on how the model performs. For this purpose, we learn word embeddings for all the vocabulary collected from the textual data by using Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which ensures that the more semantically similar the two words are, the closer the two word vectors are. For simplicity, we denote the word embeddings of the sentence $s_i$ as $s_i = \{w_1, w_2, ..., w_l\}$.

Due to the advancements achieved using convolutional neural networks (CNNs) in computer vision, image descriptors trained using CNNs over a large number of images have proven to be efficient and effective for many downstream applications. To this end, we train VGG-19 network (Simonyan & Zisserman, 2014) over the ImageNet dataset and use the pre-trained VGG-19 as image feature extractor in a transfer learning setting. Specifically, we use the output of the last fully-connected layer (FC7) before the classification layer to encode images $img$ as:

$$p = \{p_1, p_2, ..., p_m\} = \{VGG(img_1), VGG(img_2), ..., VGG(img_m)\}$$

where $p_m$ means to the visual features extracted from the $m$th image through the pretrained VGG-19.

3.4. Image-enhanced text representation learning

The most majority of fake news detection methods are text-only oriented and their great achievements have demonstrated that understanding text content semantics is importantly crucial to predict whether a news is fake or not. However, kinds of social media networks have accumulated a large quantity of multimodal news and cross-modal semantic enhancement is doomed to improve the understanding of unimodal information. Here, to enhance text representations we exploit images attached to the text to assist text representation learning and devise a hierarchical attention mechanism (Yang et al., 2016; Zeng, Yang et al., 2019) to differentiate the words of the text and the images. It can be seen from Fig. 1 that considering the hierarchical document structure we model the document hierarchically, i.e., word level (from word to sentence) and sentence level (from sentence to document).

3.4.1. Word encoder with soft attention

In word level, a Bi-directional Long-Short Term Memory (BiLSTM) network is devised to process the words of each sentence, followed by a soft attention mechanism to represent the sentence. Compared with single LSTM cell, BiLSTM cell is able to compute the sentence $s_i = \{w_1, w_2, ..., w_l\}$ from both forward and backward directions which outputs a forward hidden sequence $\{\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_l\}$ and a backward hidden sequence $\{\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_l\}$. For example, the forward hidden state $\tilde{h}_t$ in the $t$ time step can be formally denoted as:

$$\tilde{h}_t = \sigma(W_h w_t + W_i h_{t-1} + b_h)$$

$$\tilde{c}_t = \tilde{F}_t \odot \tilde{c}_{t-1} + \tilde{I}_t \odot \tanh(W_c w_t + W_i h_{t-1} + b_c)$$

$$\tilde{O}_t = \sigma(W_o w_t + W_i h_{t-1} + b_o)$$

$$\tilde{h}_t = \tilde{O}_t \odot \tanh(\tilde{c}_t)$$

$$\tilde{F}_t = \sigma(W_f w_t + W_i h_{t-1} + b_f)$$

$$\tilde{I}_t = \sigma(W_i w_t + W_i h_{t-1} + b_i)$$

$$\tilde{C}_t = \tilde{F}_t \odot \tilde{C}_{t-1} + \tilde{I}_t \odot \tanh(W_c w_t + W_i h_{t-1} + b_c)$$

$$\tilde{O}_t = \sigma(W_o w_t + W_i h_{t-1} + b_o)$$

\(\tilde{h}_t = \tilde{O}_t \odot \tanh(\tilde{c}_t)\)

\(\tilde{F}_t = \sigma(W_f w_t + W_i h_{t-1} + b_f)\)

\(\tilde{I}_t = \sigma(W_i w_t + W_i h_{t-1} + b_i)\)

\(\tilde{C}_t = \tilde{F}_t \odot \tilde{C}_{t-1} + \tilde{I}_t \odot \tanh(W_c w_t + W_i h_{t-1} + b_c)\)

\(\tilde{O}_t = \sigma(W_o w_t + W_i h_{t-1} + b_o)\)

\(\tilde{h}_t = \tilde{O}_t \odot \tanh(\tilde{c}_t)\)



Table 1

| Symbol | Description |
|--------|-------------|
| $p_m$  | The visual features extracted from the $m$th image through the pretrained VGG-19 |
| $h_t$  | The output of the $t$ time step of the word encoder in stage 2 |
| $d_t$  | The output of the hierarchical attention network |
| $R_{vis}$ | The output of the textual encoder in stage 2 |
| $R_{vis}$ | The output of the visual encoder in stage 3 |
| $s_i$  | The shared multimodal representation of the text document and the $i$th image attached to it |
where $\mathbf{\mathcal{I}}_{t}$, $\mathbf{\mathcal{I}}_{t}^{\perp}$ and $\mathbf{\mathcal{O}}_{t}$ represent the input gate, forget gate and output gate respectively. $\sigma$ is the sigmoid function controlling the in and out of information flow. $\odot$ means to be the element-wise product and others are model parameters to be learned during the training phase.

The backward hidden state $\mathbf{h}_{t}^{\perp}$ in the t time step can be computed similarly. Then the forward hidden state $\mathbf{h}_{t}^{\perp}$ and the backward hidden state $\mathbf{h}_{t}$ are combined to generate $h_t = [\mathbf{h}_{t}^{\perp}, \mathbf{h}_{t}]$ which is regarded as the output of the t time step of BiLSTM. We formally denote it as:

$$h_t = \text{BiLSTM}(s_t)$$  \hspace{1cm} (5)

Since the words are not equally important to represent the sentence, a soft attention is used as follows:

$$u_{t,i} = U \tanh(W_h h_t + b_u)$$

$$a_{t,i} = \frac{\exp(u_{t,i})}{\sum_t \exp(u_{t,i})}$$  \hspace{1cm} (4)

$$s_t = \sum_i a_{t,i} h_{t,i}$$

where $U$ is a randomly initialized context vector and learned during the training phase. We first project the hidden vector $h_t$ to an attention space through a non-linear layer of neurons with the activation function $\tanh$. Then $U$ is used as supervisor to compute the scalar $u_{t,i}$ which measures the importance of each word. Afterwards, softmax is exploited to normalize the attention weights $\{a_{t,1}, a_{t,2}, \ldots, a_{t,T}\}$. At last, we obtain the sentence representation $s_t$ via weighted summation over all its word representations. Note that the soft attention is marked in yellow in the second component of Fig. 1.

3.4.2. Sentence encoder with visual aspect attention

In sentence level, an another bi-directional Bi-directional Long-Shot Term Memory (BiLSTM) network is used to process the sentences of the document, which is followed by an attention mechanism supervised by the images attached to the document (visual aspect attention) to represent the document. For simplicity, we formally denote the BiLSTM cell operation as:

$$h_t = \text{BiLSTM}(s_t)$$  \hspace{1cm} (5)

where $h_t$ means to be the hidden vector in the t time step which is actually composed of the forward hidden vector and the backward hidden vector, i.e., $h_t = [\mathbf{h}_t^{\perp}, \mathbf{h}_t]$.

Similarly, the sentences are not equally important to represent the news document. Each image attached to the post can be used as supervisor to measure the importance of each sentence. We formally denote the visual aspect attention as follows:

$$g_{j} = \tanh(W_p p_j + b_v)$$

$$q_{j} = \tanh(W_h h_t + b_v)$$

$$v_{j,i} = V^T (g_{j} \odot q_{j} + q_{i})$$

$$\beta_{j,i} = \frac{\exp(v_{j,i})}{\sum_t \exp(v_{j,i})}$$

$$d_{j} = \sum_t \beta_{j,i} h_t$$

where $V$ is a randomly initialized context vector whose effect resembles $U$ in word level. Specifically, we first project each image representation $p_j$ and each sentence representation $h_t$ to an attention space through a non-linear layer of neurons with the activation function tanh. Then $V$ is treated as supervisor to compute the scalar $v_{j,i}$ which reveals the importance of each sentence. Afterwards, softmax is exploited to normalize the attention weights $\{\beta_{j,1}, \beta_{j,2}, \ldots, \beta_{j,T}\}$. At last, in terms of the image $p_j$ we obtain the document representation $d_j$ via weighted summation over all its sentence representations. Note that visual aspect attention is marked in pink in the second component of Fig. 1.

In order to differentiate the value of each image, we utilize a soft attention mechanism to generate the final representation of the document in document level. We formally denote the soft attention as follows:

$$k_i = K \tanh(W_d d_i + b_3)$$

$$f_i = \sum_j \exp(k_i)$$

$$d = \sum_j f_i d_i$$

where $K$ is a randomly initialized context vector whose effect resembles $U$ and $V$. Specifically, we first project each image enhanced text representation $d_i$ to an attention space through a non-linear layer of neurons with the activation function tanh. Then $K$ is treated as supervisor to compute the scalar $k_i$ which reveals the importance of each image enhanced text representation. Afterwards, softmax is exploited to normalize the attention weights $\{f_{j,1}, f_{j,2}, \ldots, f_{j,m}\}$. At last, we obtain the document representation $d$ via weighted summation over all its image enhanced text representations. Note that soft attention is marked in blue in the second component of Fig. 1.

3.5. Multimodal representation learning

In order to further capture the correlations between textual modality and visual modality, we attempt to devise a novel VAE-based multimodal feature fusion method which consists of one encoder and one decoder. We can observe from Fig. 1 that the encoder can be divided into two sub-components: textual encoder and visual encoder, and correspondingly the decoder can be divided into two sub-components as well: textual decoder and visual decoder. The four modules are marked in orange, grey, green and blue in the third component of Fig. 1 respectively. We will elaborate them in the following.

3.5.1. Textual encoder

Given a text document $\text{doc}$, the textual encoder treats it as a long sequence which has $T$ words, $\text{doc} = \{w_1, w_2, \ldots, w_T\}$. Each word $w_t$ represents a word vector learned by the first module. In order to extract textual semantics, we utilize a stacked Bi-directional Long-Shot Term Memory (BiLSTM) network to model the text document. Specifically, the stacked BiLSTM consists of two Bi-LSTM and one full connected layer and is denoted as:

$$H_t = \text{BiLSTM}(w_t)$$

$$H_{\text{txt}} = \text{BiLSTM}(H_t)$$

$$R_{\text{txt}} = \tanh(W_{\text{txt}} H + b_{\text{txt}})$$

where $H = \{H_1, H_2, \ldots, H_T\}$ is the output of the second BiLSTM, and $R_{\text{txt}}$ denotes text feature extracted through a fully connected layer.

3.5.2. Visual encoder

Images $\text{imgs}$ attached to a text document $\text{doc}$ are vectorized as $p = \{p_1, p_2, \ldots, p_m\}$, followed by two full connected layers. We formally denote the visual encoder as:

$$\tilde{R}_{\text{vis}}^{\text{in}} = \tanh(W_{\text{vis}} p_t + b_{\text{vis}})$$

$$R_{\text{vis}}^{\text{in}} = \tanh(W_{\text{vis}} \tilde{R}_{\text{vis}}^{\text{in}} + b_{\text{vis}})$$

where $R_{\text{vis}}^{\text{in}}$ denotes the visual feature of the image $\text{img}_t$ extracted through two fully connected layers.
After the text feature \( R_{tx} \) and each image feature \( R_{tx}^i \) are produced, we first concatenate them and project it into a latent space through a fully connected layer as follows.

\[
R_{tx}^i = \tanh(W_{tx}[R_{tx}, R_{tx}^i] + b_{tx})
\]  

(10)

Then we compute the mean \( \mu \) and the variance \( \sigma \) from the latent space as follows.

\[
z_i = \mu + \sigma \odot \epsilon
\]  

(11)

where \( z_i \) denotes the shared multimodal representation of the text document and the \( i \)th image attached to it, and \( \epsilon \) is a random variable sampled from a Gaussian distribution.

To sum up, aiming at mapping the multimodal data into a latent space, the encoder takes as input the text document and images attached to it, and outputs shared multimodal representations. For simplicity, we formally denote it as:

\[
z_i = \text{Encoder}(\text{doc, img}_i)
\]  

(12)

3.5.3. Textual decoder

We can clearly see from the architecture of our proposed method that the decoder is structurally similar to the encoder. In other words, the encoder and the decoder are inverted from the perspective of network structure. The textual decoder aims to reconstruct the words in the text according to the multimodal representations. To this end, we first use a fully connected layer to process the multimodal representations. Then a stacked Bi-LSTM is leveraged to generate words during all time steps.

3.5.4. Visual decoder

The visual decoder aims to reconstruct the visual features extracted by the pretrained VGG-19 model according to the multimodal representations. To this end, we first use a fully connected layer to process the multimodal representations. Then a stacked Bi-LSTM is leveraged to generate words during all time steps.

3.6. Fake news detector

In this work, we detect fake news through learning the multimodal fusion eigenvector and image-enhanced text representation in a multitask learning setting. In order to obtain the multimodal fusion eigenvector, we exploit a soft attention to differentiate the importance of each image. We formally denote it as follows:

\[
k_i = Q^T \tanh(W_i z_i + b_i)
\]

\[
\chi_i = \frac{\exp(k_i)}{\sum_j \exp(k_j)}
\]

\[
z_i = \sum_i \chi_i z_i
\]  

(14)

where \( Q \) is a randomly initialized context vector whose effect resembles \( U \), \( V \) and \( K \).

Then the multimodal fusion eigenvector and the image-enhanced text representation are concatenated and processed through two fully connected layers. Finally, a Softmax based classifier is used to compute the probability of this news being fake.

\[
F = \tanh(W_{txt}[\hat{d}, \hat{z}; b_{txt}])
\]

\[
\rho = \tanh(W_{vis}^2 F + b_{vis}^2)
\]

\[
r_c = \sum \exp(\rho_c)
\]  

(15)

where \( g_c \) is the predicted probability distribution.

3.7. Model training

In this section, we need to optimize all the parameters notated as \( \Theta = \{\Theta_1, \Theta_2, \Theta_3\} \) where \( \Theta_1, \Theta_2, \Theta_3 \) represent the model parameters of image-enhanced representation learning module, multimodal representation learning module and fake news detector module respectively. First, a VAE loss is designed for multimodal representation learning module. The VAE loss includes three parts, i.e., text reconstruction loss, VGG-19 features reconstruction loss and the KL divergence loss that ensures the probability distribution to become approximately subordinate to Normal distribution. We formally denote the VAE loss as follows:

\[
L_{\text{enc}} = -\sum_{i \in D} T \sum_{c=1}^C \sum_{j=1}^{n_c} 1_{y_i=c} \log \hat{g}_c^j
\]  

(16)

\[
L_{\text{rec}} = \frac{1}{n_i} \sum_{i \in D} \sum_{j=1}^V (\hat{y}_i^j - y_i^j)^2
\]  

(17)

\[
L_{\text{4l}} = \frac{1}{2} \sum_{i \in D} (\hat{p}_i^j + c_i^j - \log(\sigma_i) - 1))
\]  

(18)

where \( T \) is the number of words in one text document, \( n_v \) is the vocabulary size, \( n_c \) is the dimensionality size of VGG-19 features, and \( n_m \) is the dimensionality size of multimodal features. \( \hat{c}_i \) means to be the \( i \)th word in the text document. \( \hat{c}_i \) is the reconstructed word. \( y_i^j \) is the \( i \)th word visual features extracted by pretrained VGG-19 model, and \( \hat{y}_i^j \) is the reconstructed one.

Second, cross entropy with \( L_2 \) regularization is defined as the loss function of image-enhanced representation learning module and fake news detector when training:

\[
L_{\text{ind}} = -\sum_{x \in C} y_{c(x)} \log(g_c(x)) + \lambda L_2(\Theta_1, \Theta_2)
\]  

(19)

where \( C \) is the number of categories, \( y_{c(x)} \) is the golden truth, \( g_c(x) \) is the predicted probability and \( \lambda \) is the coefficient for \( L_2 \) regularization.

The VAE loss and the cross entropy loss are optimized using SGD to deeply mine the correlations between textual modality and visual modality for fake news detection.

4. Experiments

In this section, we first describe the experimental settings in brief (Section 4.1). Then, four evaluation metrics are introduced in Section 4.2. Afterwards, several baselines are listed in Section 4.3. At last, in order to evaluate the proposed FND-SCTI, comparison experiments are conducted and empirical results are analyzed in Section 4.4.

4.1. Experimental settings

We evaluate the proposed FND-SCTI on two real-world multimodal datasets, i.e., Twitter (Boididou et al., 2015) and Weibo (Jin et al., 2017). The statistics of the used two real-world datasets are summarized in Table 2. The Twitter dataset contains 7898 fake news, 6026 real news and 514 images, respectively. The Weibo dataset contains 4749 fake news and 514 images.

Table 2

| Dataset | Fake news | Real news | Images |
|---------|-----------|-----------|--------|
| Twitter | 7898      | 6026      | 514    |
| Weibo   | 4749      | 4779      | 9528   |
news, 4779 real news and 9528 images, respectively. We preprocess the Twitter dataset by filtering out tweets attached with videos as done in Dhruv et al. (2019). As for the Weibo dataset which was released by an authoritative China news agency and verified by Weibo’s official rumor debunking system, we preprocess it by removing duplicate images as done in Dhruv et al. (2019). Each dataset is split into training and testing set.

In our experiments, we pre-train the Word2Vec model (Mikolov et al., 2013) on the dataset unsupervisedly and generate word embeddings with dimension size of 32 for all the vocabulary. As for visual feature extractor, we pre-train a VGG-19 model (Simonyan & Zisserman, 2014) on ImageNet dataset and use the output of the second to the last layer as visual features with dimension size of 4096. The parameters of pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are.

In our experiments, we pre-train the Word2Vec model (Mikolov et al., 2013) on the dataset unsupervisedly and generate word embeddings with dimension size of 32 for all the vocabulary. As for visual feature extractor, we pre-train a VGG-19 model (Simonyan & Zisserman, 2014) on ImageNet dataset and use the output of the second to the last layer as visual features with dimension size of 4096. The parameters of pre-trained VGG-19 model are frozen during training. All LSTM cells and fully connected layers have dimension size of 32. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The two fully connected layers of the textual encoder have dimension size of 1024 and 32 respectively. The textual decoder and the visual decoder share the same dimensions with the textual encoder and the visual encoder. Context vectors are.

4.2. Evaluation metrics

Here, we use four evaluation metrics, i.e., classification accuracy, precision, recall and F1-score, which are widely used for fake news detection. Accuracy measures how the method performs in classification, and is formalized as:

\[
\text{Accuracy} = \frac{N_{\text{correct}}}{N} \quad (20)
\]

where \(N_{\text{correct}}\) is the number of samples correctly predicted and \(N\) is the total number of the tested news.

Precision measures the probability of true positive samples over all the predicted positive samples, and is formalized as:

\[
\text{Precision} = \frac{N_p}{N_p + N_f} \quad (21)
\]

where \(N_p\) represents the number of positive samples which are predicted to be positive, and \(N_f\) represents the number of negative samples which are predicted to be positive.

Recall measures the probability of true positive samples over all the original positive samples, and is formalized as:

\[
\text{Recall} = \frac{N_p}{N_p + N_n} \quad (22)
\]

where \(N_n\) represents the number of positive samples which are predicted to be false.

F1-score is the harmonic mean of the precision and the recall, and can be formulated as:

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (23)
\]

### 4.3. Baseline approaches

In order to comprehensively validate our proposed FND-SCTI, we select several competitive fake news detection algorithms for a fair comparison. Two unimodal approaches (Textual and Visual) and five multimodal approaches (VQA, Neural Talk, att-RNN, EANN and MVAE) are listed as follows:

- **Textual** (Dhruv et al., 2019) aims to use a Bi-LSTM to model the text-only content of the news, followed by a 32-dimensional fully connected layer. Then a softmax layer is leveraged for final classification.

- **Visual** (Dhruv et al., 2019) aims to model the images attached to news posts with a pre-trained VGG-19, followed by two fully connected layers with dimension of 1024 and 32 respectively.

- **VQA** (Antol et al., 2015) targets at discovering clues from the images and answering questions. The original VQA model is adapted to multimodal fake news detection by transforming the multi-class classification to a binary-class classification and setting the dimension of LSTM cell to be 32.

- **Neural Talk** (Vinyals et al., 2015) is proposed to deal with image captioning which produces sentences to depict images. To adapt Neural Talk to fake news detection, we average the output of the LSTM hidden values as the multimodal features which are fed into a fully connected layer.

- **att-RNN** (Jin et al., 2017) is designed for rumor detection, which exploits a neural attention on top of the LSTM to fuse text features, social context and image features. We adapt it by removing social context to ensure a fair comparison.

- **EANN** (Event Adversarial Neural Network) (Wang et al., 2018) is proposed to extract event-agnostic features for multimodal fake news
4.4. Experimental results and analysis

In order to demonstrate that our proposed FND-SCTI is able to perform better for multimodal fake news detection via thoroughly mining the semantic correlations between textual modality and visual modality, we make comparisons with two unimodal methods and five multimodal methods. The qualitative comparison results are shown in Table 3. Note that the results of the competitors on the two datasets stems from Dhruv et al. (2019) and they all regard one text document attached with multiple images as multiple instances. Therefore, we conduct experiments in this same experiment setting for fair comparisons.

First, it can be clearly observed from the results that multimodal methods perform much better than unimodal methods, which shows the necessity to utilize multiple modalities. Second, comparing at – RNN with VQA and Neural Talk, we find that visual aspect attention is of great importance to improve the performance, which validate the necessity to exploit images attached to enhance text representation. Third, comparing MVAE with VQA, Neural Talk, at – RNN and EANN, we can conclude that learning the shared multimodal representations for textual modality and visual modality contributes largely to detecting fake news, which validate the necessity to fuse multimodal features. Finally, the reason why our proposed FND – SCTI wins the current state-of-the-art approach MVAE is that FND – SCTI learns the multimodal fusion representation and the image-augmented text representation in a multi-task learning setting. In other words, the deeper the semantic correlations between textual modality and visual modality are mined, the better the method performs. To be specific, in terms of detection accuracy, FND – SCTI outperforms MVAE with absolute increments of 1.3% and 1.0% on the Twitter dataset and Weibo dataset respectively.

We further conduct an ablation experimental analysis to investigate the respective contributions of the components of FND – SCTI’s architecture. Note that we regard one text document attached with multiple images as one instance. There are two reasons for this experimental setting: (1) This setting is much more practical and reasonable than that one text document attached with multiple images are treated as multiple instances; (2) The visual aspect attention does not work when one text document attached with multiple images are treated as multiple instances, that is to say, this setting ensures visual aspect attention to make a big contribution.

Table 4 displays the results of the ablation analysis. FND – SCTI – han removes the multimodal representation learning module while FND – SCTI – visatt removes the image-enhance text representation learning module. We can clearly see from the results that: (1) FND – SCTI outperforms FND – SCTI – han with absolute increments of 8.9% and 6.5% on average on the Twitter dataset and Weibo dataset respectively, which directly validate the efficacy of our devised multimodal representation learning module; (2) FND – SCTI outperforms FND – SCTI – visatt with absolute increments of 1.4% and 1.0% on average on the Twitter dataset and Weibo dataset respectively, which directly validate the efficacy of our devised image-enhanced text representation learning module.

All the substantial experimental results show that our proposed FND – SCTI is effective in detecting fake news on social medial networks and achieves the new state-of-the-art performance.

5. Conclusion and future work

In this paper, we are targeting at detecting whether a multimodal news is fake or not by analyzing both textual and visual contents. We present a novel deep neural network based method, called FND-SCTI which not only learns a shared multimodal fusion representation, but also enhances textual representation by devising hierarchical attention to highlight the important part of news document. The fake news detector takes as input the two features learned above and is trained in a multi-task learning setting. Extensive experiments conducted on Twitter and Weibo datasets have demonstrated that our proposed FND-SCTI is able to effectively capture the semantic correlations across multimodalities, and achieves the state-of-the-art performance.

Although it is empirically validated that our proposal has shown great potentials for multimodal fake news detection, current deep learning based methods leave space for improvement. Both domestic and foreign social networks are flooded with bot accounts which are mainly used to propagate advertisements or false information. In the future, we plan to do source checking by modeling user characteristics. With the development of image editing software like Photoshop and deep generative networks like GANs, image editing and generation becomes increasingly simpler to operate, which makes it more and more difficult to distinguish fake images from real ones. Therefore, forged image detection is our another plan.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful and thought-provoking suggestions and feedbacks. This work was supported in part by the National Natural Science Foundation of China under grant No. 61802440, the Natural Science Foundation of Hubei Province under grant No. 2020CFB492 and the Basic Scientific Research of China University under grant No. 30106200278.

References

Al-Turjman, F., & Deebak, B. D. (2020). Privacy-aware energy-efficient framework using the Internet of medical things for covid-19. IEEE Internet of Things Magazine, 3(3), 64-68.
Lu, Y.-J., & Li, C.-T. (2020). Geant: Graph-aware co-attention networks for explainable fake news detection on social media (arXiv preprint) arXiv:2004.11648.

Ma, J., Gao, W., & Wong, K.-F. (2018a). Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics, 171–182.

Majumdar, R., & Singhal, A. (2019). Challenges and opportunities in fake news detection on social media. IEEE Access, 7, 163608–163617.

Ma, J., Gao, W., & Wong, K.-F. (2018b). Detect rumor and stance jointly by neural multi-task learning. companion proceedings of the – The web conference 2018, 585–593.

Ma, X., Zeng, J., Feng, L., Fortino, G., & Zhang, Y. (2019). Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. Future Generation Computer Systems, 104, 304–311.

Masoud, R. A., & Mirmahaleh, S. Y. H. (2020). Coronavirus disease (covid-19) prevention and treatment methods and effective parameters: A systematic literature review. Sustainable Cities and Society, 57, 102071.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013, 3111–3119.

Qi, F., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. In 2019 IEEE international conference on data mining (ICDM) (pp. 518–527).

Shu, K., Wang, S., & Liu, H. (2019a). Beyond news contents: The role of social context for fake news detection. Proceedings of the twelfth ACM international conference on web search and data mining, 312–320.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019b). Explainable fake news detection. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 395–405.

Shui-Hua, W., & Zhang, Y.-D. (2020). DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16, 1–19.

Simonony, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition (arXiv preprint) arXiv:1409.1556.

Song, Z., Srivastava, S., & Al-Turjman, F. (2020). A systematic approach for covid-19 predictions and parameter estimation. Personal and Ubiquitous Computing, 1–13.

Truong, Q.-T., & Lauw, H. W. (2019). Vistanet: Visual aspect attention network for multimodal sentiment analysis. Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 305–312.

Vinyls, O., Toshnev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE conference on computer vision and pattern recognition, 2249–2257.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). Eann: Event adversarial autoencoder for fake news detection. In Proceedings of the 34th AAAI conference on artificial intelligence. AAAI 2020, the thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the 34th AAAI conference on artificial intelligence, March 2-7, 2020,held as an online conference, 2020 (pp. 516–520).

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by composite learning factor and precomputation for multiple sclerosis classification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16, 1–19.

Zeng, J., Ma, X., & Zhou, K. (2019). Photo-realistic face age progression/regression using deep learning. IEEE Access, 7, 2921–2934.

Zeng, J., Gao, W., & Wong, K.-F. (2018a). Graph-aware co-attention networks for explainable fake news detection on social media (arXiv preprint) arXiv:1409.0473.

Zeng, J., Gao, W., & Wong, K.-F. (2018b). Detect rumor and stance jointly by neural multi-task learning. Companion proceedings of the – The web conference 2018, 585–593.