Information Processing and Management 58 (2021) 102437

Contents lists available at ScienceDirect

Information Processing and Management

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

A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks

Chenguang Song [a], Nianwen Ning [a], Yunlei Zhang [b], Bin Wu [*,a]

[a] Beijing Key Laboratory of Intelligence Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] North China Institute of Science and Technology, Hebei 065201, PR China

ARTICLE INFO

Keywords:
Fake news detection
Crossmodal attention
Residual network
Convolutional neural network

ABSTRACT

In recent years, social media has increasingly become one of the popular ways for people to consume news. As proliferation of fake news on social media has the negative impacts on individuals and society, automatic fake news detection has been explored by different research communities for combating fake news. With the development of multimedia technology, there is a phenomenon that cannot be ignored is that more and more social media news contains information with different modalities, e.g., texts, pictures and videos. The multiple information modalities show more evidence of the happening of news events and present new opportunities to detect features in fake news. First, for multimodal fake news detection task, it is a challenge of keeping the unique properties for each modality while fusing the relevant information between different modalities. Second, for some news, the information fusion between different modalities may produce the noise information which affects model's performance. Unfortunately, existing methods fail to handle these challenges. To address these problems, we propose a multimodal fake news detection framework based on **C**rossmodal **A**ttention **R**esidual and **M**ultichannel convolutional neural **N**etworks (CARMN). The **C**rossmodal **A**ttention **R**esidual **N**etwork (CARN) can selectively extract the relevant information related to a target modality from another source modality while maintaining the unique information of the target modality. The **M**ultichannel **C**onvolutional neural **N**etwork (MCN) can mitigate the influence of noise information which may be generated by crossmodal fusion component by extracting textual feature representation from original and fused textual information simultaneously. We conduct extensive experiments on four real-world datasets and demonstrate that the proposed model outperforms the state-of-the-art methods and learns more discriminable feature representations.

1. Introduction

As its low cost, convenience, and rapid propagation of information, social media has gradually become one of the important platforms for people to seek out and consume news in recent years Shu, Cui, Wang, Lee, and Liu (2019); Shu, Sliva, Wang, Tang, and Liu (2017); Zhang and Ghorbani (2020). Compared with traditional news media, the lack of effective supervision measures for social

* Corresponding author.
E-mail address: wubin@bupt.edu.cn (B. Wu).

https://doi.org/10.1016/j.ipm.2020.102437
Received 29 February 2020; Received in revised form 27 September 2020; Accepted 8 November 2020
Available online 16 November 2020

media weakens the journalistic norms of objectivity. One can publish and spread fake news on social media at a very low cost. The proliferation of fake news on social media will bring negative impacts on both individuals and society. It may undermine the traditional news sources that have enjoyed high levels of public trust and credibility and harm stability and harmony of society Lazer et al. (2018).

One of the methods to mitigate the serious negative effects caused by the fake news is manual fact-checking Zhou, Zafarani, Shu, and Liu (2019b), which includes expert-based fact-checking and crowd-sourced manual fact-checking. Expert-based fact-checking can obtain high accuracy but needs intensive labor and cost of time and has difficulty in scaling with emerging fake news. Crowd-sourced manual fact-checking do well in scalability but will get a relatively less credible label, which fails to meet the qualification of accurate fake news detection. As the limitations of manual fact-checking approaches, automatic fake news detection techniques have been developed to solve the problem Zhou et al. (2019b). Some early researchers try to manually design a series of features which are fed into a machine learning model to identify fake news Castillo, Mendoza, and Poblete (2011); Sejeong, Meeyoung, Kyomin, Wei, and Yajun (2013); Yang, Liu, Yu, and Yang (2012), but these methods are still time-consuming and poor in generalizability.

As the powerful ability of the deep neural networks (DNN) to automatic capture complex patterns, it was introduced to alleviate the shortcomings of traditional methods. Most of existing studies are mainly focus on using the textual features to detect fake news Oshikawa, Qian, and Wang (2018); Su, Macdonald, and Ounis (2019). However, there is a phenomenon that cannot be ignored is that more and more social media news contains information with different modalities such as texts, pictures, and videos. There are complementary and enhanced relationships between different modalities Cao et al. (2018); Cui, Wang, and Lee (2019); Zhao et al. (2019). More importantly, news with visual information is likely to attract much more attention from users and thus gains a larger propagation range Jin, Cao, Guo, Zhang, and Luo (2017a); Qi, Cao, Yang, Guo, and Li (2019). But limited work has been performed on verifying the credibility of news by exploiting visual information. Jin et al. first proposed a Recurrent Neural Networks (RNN)-based automatic multimodal fake news detection model, in which the multimodal features are fused via attention mechanism Jin et al. (2017a). Wang et al. proposed a multi-task learning framework to learn both textual and visual transferable feature representations among all the posts by leveraging an additional event discriminator Wang et al. (2018b). A similar idea is that Zhang et al. proposed an event memory network module to learn invariant features among different events Zhang, Fang, Qian, and Xu (2019). Khattar et al. proposed a multimodal fusion fake news detection framework based on Variational Autoencoder (VAE) Khattar, Goud, Gupta, and Varma (2019).

First, despite great progress has been made in previous research, an important problem is ignored—how to keep the unique properties for each modality while fusing the relevant information between different modalities. Textual and visual feature representations are learned by different ways and should have their own unique characteristic. It is not a good choice to fuse different modal feature representations to one. Different modal feature separately representations will fail to fuse the correlative and complementary information between different modalities. Second, it should be noted that multimodal fake news detection task usually uses high-level image embeddings and low-level sentence embeddings Jin et al. (2017a) and the visual feature representation is extracted from the model pretrained on Imagenet set Simonyan and Zisserman (2015), which means that it is impossible to accurately match text and image information. And fake news images from social media have more complex patterns at both physical and semantic levels Cao et al. (2020); Qi et al. (2019). For some posts and their attached images, the visual feature representations extracted from pretrained model are not always what we expect. The fusion between textual and visual information may produce noise information which may affect model's performance. Thus, we should consider both original and fused text information simultaneously. Existing multimodal fake news detection methods fail to meet these requirements.

To overcome the limitations of existing approaches, a multimodal fake news detection model based on **C**rossmodal **A**ttention **R**esidual and **M**ultichannel convolutional neural **N**etworks (CARMN) is proposed in this paper. The **C**rossmodal **A**ttention **R**esidual **N**etwork (CARN) can selectively extract the information related to a target modality from another source modality while maintaining the unique information of the target modality. The **M**ultichannel **C**onvolutional neural **N**etwork (MCN) can extract textual feature representation from original and fused textual information simultaneously and mitigate the influence of noise information which may be generated by crossmodal fusion component Wang, Zhang, Xie, and Guo (2018a). At present, there are only a few reliable social-media-oriented multimodal fake news detection datasets. Thus, we collected a large number of reliable fake and real news from the Weibo platform[1]. Our main contributions can be summarized as follows.

- We present a novel multimodal fake news detection model based on CARN and MCN.
- The CARN is introduced to fuse the relevant information between different modalities and keep the unique properties for each modality.
- To mitigate the influence of noise information which may be generated by crossmodal fusion, the MCN is introduced to extract feature representations from original and fused textual information simultaneously.
- We conduct extensive experiments on four real-world datasets and demonstrate that the proposed model outperforms state-of-the-art methods and learn more discriminable feature representations.
- We contribute a large scale multimodal fake news dataset from Weibo platform and will make it available to the public[2].

[1] http://www.weibo.com/
[2] https://github.com/lumen2018/dataset

2. Related work

2.1. The definition of fake news

Fake news overlaps with other concepts and terms such as false news, rumor, and disinformation Ajao, Bhowmik, and Zargari (2018); Bondielli and Marcelloni (2019); Lazer et al. (2018). An universal definition for fake news is still missing so far Zhou et al. (2019b). Similar to previous work Khattar et al. (2019); Wang et al. (2018b), we define fake news to be verifiable false news.

2.2. Fake news detection

The fake news detection task. The fake news detection task aims to assess the authenticity for the given news Kakol, Nielek, and Wierzbicki (2017); Zhou et al. (2019b). Most existing approaches formulate the fake news detection problem as a binary classification problem (fake or real) Shu et al. (2017). In some cases, it also is considered as multi-classification Karimi, Roy, Saba-Sadiya, and Tang (2018), regression, or clustering problems Oshikawa et al. (2018). The way of binary classification is adopted in this paper. The literature on fake news detection is extensive. We will provide a brief review of the work from the following categories: text-based, user-based, propagation-based and multimodal fake news detection.

Text-based fake news detection. The early studies obtain text-based features by manual linguistic cues selection Rubin, Chen, and Conroy (2015); Ruchansky, Seo, and Liu (2017). For fake news detection task, it is difficult to generalize hand-crafted linguistic features across topics and domains Sharma et al. (2019). A RNN-based model was introduced to automatically learn the hidden representation of temporal textual feature, which outperforms the methods leveraging hand-crafted features Ma et al. (2016). In order to capture the long-range dependency among variable length sequential information, Chen et al. adopted soft-attention and RNN to learn selectively temporal feature representation of post series Chen, Li, Yin, and Zhang (2018). Similarly, Yu et al. proposed a convolutional neural networks (CNN)-based model which is used to extract low-level local-global features from the input sequences and then construct high-level interactions among important features Yu, Liu, Wu, Wang, and Tan (2017). By exploiting the users' feedback towards a target claim, stance information was proved to be a strong indicator for classification Dungs, Aker, Fuhr, and Bontcheva (2018); Kochkina, Liakata, and Zubiaga (2018); Ma, Gao, and Wong (2018a), but each response has to be given a special stance label, which is laborious. Inspired by Generative Adversarial Networks (GAN), Ma et al. proposed a GAN-style fake news detection model Ma, Gao, and Wong (2019). Textual feature representation is improved by adversarial learning between text generator and fake news discriminator. Scholars also explored text-based fake news detection with various way such as user response generating Qian, Gong, Sharma, and Liu (2018), text generation Vo and Lee (2019), reinforcement learning Zhou, Shu, Li, and Lau (2019a), fact-checking url recommendation Vo and Lee (2018) and attention-residual network Chen, Sui, Hu, and Gong (2019).

User-based and propagation-based fake news detection. Apart from textual features, user profiles and propagation-based features as auxiliary information are also used to help differentiate fake news. Shu et al. provided a systematic research about the relationship between user profiles and the credibility of news Shu, Wang, and Liu (2018). Guo et al. fused the propagation features and user profiles with textual features via attention mechanism Guo, Cao, Zhang, Guo, and Li (2018). In addition, diffusion-based models have been introduced to solve this problem. Vosoughi et al. claimed that fake news tend to spreads faster, farther and more broadly than the truth on social network Vosoughi, Roy, and Aral (2018). According to supporting and opposing relations among posts, Jin et al. designed a homogeneous stance signed network to evaluate news credibility Jin, Cao, Zhang, and Luo (2016). Similarly, by exploiting post-repost relationships, Ma et al. proposed two kinds of recursive neural network models based on bottom-up and top-down tree-structured Ma, Gao, and Wong (2018b).

Multimodal fake news detection. Different from all the aforementioned work, visual information, as auxiliary information, also has been adopted to infer the veracity of news articles Gupta, Lamba, Kumaraguru, and Joshi (2013); Gupta, Zhao, and Han (2012); Ke, Song, and Kenny Q (2015). There only a few studies that focus on the correlation between image and credibility of tweets Cao et al. (2018). By introducing some features from the field of image retrieval, Jin et al. first provided a systematic research on image features between fake and real news Jin, Cao, Zhang, Zhou, and Tian (2017b). However, these features are still hand-crafted and do not capture the complex visual content information. Inspired by DNN that achieved impressive results for image and text feature representation task, Jin et al. proposed a RNN-based multimodal fusion fake news detection framework Jin et al. (2017a). The high-level visual features and high-level textual and social features are fused by attention mechanism. Wang et al. proposed a multi-task learning model to learn textual and visual transferable feature representations among all the posts by removing textual and visual event-specific information Wang et al. (2018b). Similarly, for the event-level fake news detection task, Zhang et al. used the memory network to learn event invariant features and obtained better generalizability for newly emerged events Zhang et al. (2019). In order to learn a shared latent representation across modalities, Khattar et al. proposed a multi-modal fusion framework based on VAE Khattar et al. (2019). Recently, transfer learning-based methods also have been introduced to verify the authenticity of news Singhal et al. (2020); Singhal, Shah, Chakraborty, Kumaraguru, and Satoh (2019).

3. Problem formulation

There are two ways to detect fake news: post-level or tweet-level (to identify a single post is fake/real news) and event-level (to identify a news which include a group of posts is fake/real). Our research falls in the former. Let T be a post, $T = [w_1, w_2, \cdots w_n]$, where n is the number of words w and P is an attached image of the post T. Given a post T and an attached image P, the task of this paper is to

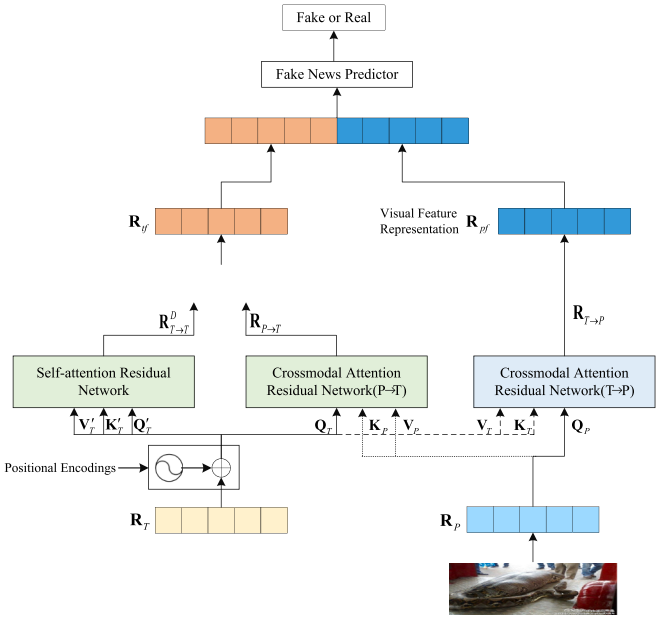

Fig. 1. The proposed model framework.

identify the post T is real ($y = 0$) or fake ($y = 1$) news by learning a fake news detection function $F : F(T, P) \rightarrow (\hat{y})$.

4. Model

4.1. Model framework

The overall structure of the proposed model is shown in Fig. 1. Our model consists of: (1) input embedding layer (to get word embedding matrix $\mathbf{R}_T$ and image embedding matrix $\mathbf{R}_P$), (2) CARN layer (to reinforce the target modality feature representation by selectively extracting information from another source modality) and self-attention residual network layer (to capture the interactions between different sequence element pairs and transmit original textual information to MCN), (3) MCN layer (to alleviate the effect of noise information which may be generated by CARN layer and extract the final textual feature representation $\mathbf{R}_{tf}$), (4) fake news prediction component (to predict a post is real or fake news). Next, we will present the details of the proposed fake news detection model.

4.2. Input embeddings

4.2.1. Word-Level sentence embeddings

For a sentence $T = \{w_1, w_2, \cdots w_n\}$, each w_i represent $i - th$ word of the sentence T and n is the length of the sentence T. Then, we convert each word of sentence T to a pretrained word embedding $\mathbf{e}_i$:

$$\mathbf{e}_i = \text{WordEmbed}(w_i) \tag{1}$$

where $\mathbf{e}_i \in \mathbb{R}^{d_T}$, d_T is is the dimension of word embeddings. The word-level sentence embeddings (i.e., word embedding matrix) of the sentence T can be denoted as:

$$\mathbf{R}_T = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\} \tag{2}$$

where $\mathbf{R}_T \in \mathbb{R}^{L_T \times d_T}$, L_T equals the length of the sentence T.

4.2.2. Image embeddings

Given m attached images $P = \{p_1, p_2, \cdots p_m\}$ of the sentence T, we extract the initial image embeddings $\mathbf{R}_{vgg}$ from the VGG-19 net pretrained on Imagenet set Simonyan and Zisserman (2015), which followed by a fully connected layer to transform initial image embeddings $\mathbf{R}_{vgg}$ to the image embedding matrix $\mathbf{R}_P$ with same dimension of word embeddings. Note that $m = 1$ in this paper. The image embedding matrix $\mathbf{R}_P$ of the sentence T can be denoted as:

$$\mathbf{R}_P = \sigma\left(\mathbf{R}_{vgg} \times \mathbf{W}_{bf} + b_P\right) \tag{3}$$

where $\mathbf{R}_{vgg} \in \mathbb{R}^{L_P \times d_{vgg}}$, $\mathbf{R}_P \in \mathbb{R}^{L_P \times d_P}$, $\mathbf{W}_{bf} \in \mathbb{R}^{d_{vgg} \times d_P}$ is the weight matrix of the fully connected layer, b_P is bias term, L_P equals to the number of the attached images m, d_{vgg} denotes the output embedding dimension of the VGG-19 ($d_{vgg} = 1000$), d_P represents the dimension of image embeddings and σ is the Leaky ReLU activation function. It should be noted that $d_P = d_T$ and $L_P = m = 1$ in this paper.

4.3. Attention residual network

4.3.1. Positional encoding

Compared to RNN, the attention-based neural networks can improve the training speed and capture longer dependencies in a sentence. However, there is no essential position information in an attention-based network. In order to enable each token of a sequence carries unique position information, the sentence embeddings are added with the positional encoding Vaswani et al. (2017). Given the embedding matrix of sentence T $\mathbf{R}_T \in \mathbb{R}^{L_T \times d_T}$, its positional encoding (PE) can be computed by:

$$\mathrm{PE}(pos, 2j) = \sin\left(pos \,/\, 10000^{2j/d_T}\right) \tag{4}$$

$$\mathrm{PE}(pos, 2j+1) = \cos\left(pos \,/\, 10000^{2j/d_T}\right) \tag{5}$$

where $pos \in [0, \cdots, L_T]$ is position, $j \in [0, d_T/2)$ is the dimension. Each dimension of the PE corresponds to a sinusoid. Then, the position information is added to a sentence representation by summing token embeddings and corresponding PE (i.e., $\mathbf{R}_T + \mathrm{PE}(\mathbf{R}_T)$).

4.3.2. Crossmodal and unimodal attention

As scaled dot-product attention is the core component of our model, we will provide the definitions of single head crossmodal attention, single head unimodal self-attention and multi-head unimodal self-attention, respectively Tsai et al. (2019); Vaswani et al. (2017). The task of crossmodal attention is to capture the relevant and complementary information between textual and visual information. When pass information from the sentence T to its attached image P (i.e., $T \rightarrow P$), the Queries, Keys and Values are defined as $\mathbf{Q}_P = \mathbf{R}_P \times \mathbf{W}_{Q_P}$, $\mathbf{K}_T = \mathbf{R}_T \times \mathbf{W}_{K_T}$ and $\mathbf{V}_T = \mathbf{R}_T \times \mathbf{W}_{V_T}$, where $\mathbf{W}_{Q_P} \in \mathbb{R}^{d_P \times d_k}$, $\mathbf{W}_{K_T} \in \mathbb{R}^{d_T \times d_k}$ and $\mathbf{W}_{V_T} \in \mathbb{R}^{d_T \times d_v}$. Note that $d_k = d_v = d_T$. The single head crossmodal attention function $Att_{T \rightarrow P} \in \mathbb{R}^{L_P \times d_v}$ is defined as follows.

$$\begin{aligned} Att_{T \rightarrow P} &= \mathrm{softmax}\left(\mathbf{Q}_P \times \mathbf{K}_T^{\top} \Big/ \sqrt{d_k}\right) \times \mathbf{V}_T \\ &= \mathrm{softmax}\left(\mathbf{R}_P \times \mathbf{W}_{Q_P} \times \mathbf{W}_{K_T}^{\top} \times \mathbf{R}_T^{\top} \Big/ \sqrt{d_k}\right) \times \mathbf{R}_T \times \mathbf{W}_{V_T} \end{aligned} \tag{6}$$

when the information from modality P is passed to modality T:

$$Att_{P \rightarrow T} = \left[\mathrm{softmax}\left(\mathbf{Q}_T \times \mathbf{K}_P^{\top} \Big/ \sqrt{d_k}\right)^{\top}\right]^{\top} \times \mathbf{V}_P \tag{7}$$

where $\mathbf{Q}_T = \mathbf{R}_T \times \mathbf{W}_{Q_T}$, $\mathbf{K}_P = \mathbf{R}_P \times \mathbf{W}_{K_P}$, $\mathbf{V}_P = \mathbf{R}_P \times \mathbf{W}_{V_P}$, $\mathbf{W}_{Q_T} \in \mathbb{R}^{d_T \times d_k}$, $\mathbf{W}_{K_P} \in \mathbb{R}^{d_P \times d_k}$, $\mathbf{W}_{V_P} \in \mathbb{R}^{d_P \times d_v}$ and $Att_{P \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$. Similarly, the single head unimodal self-attention function $Att_{T \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$ can be represented as:

$$Att_{T \rightarrow T} = \mathrm{softmax}\left[\mathbf{Q}'_T \times (\mathbf{K}'_T)^{\top} \Big/ \sqrt{d_k}\right] \times \mathbf{V}'_T \tag{8}$$

where $\mathbf{Q}'_T = \mathbf{R}_T \times \mathbf{W}'_{Q_T}$, $\mathbf{K}'_T = \mathbf{R}_T \times \mathbf{W}'_{K_T}$, $\mathbf{V}'_T = \mathbf{R}_T \times \mathbf{W}'_{V_T}$, $\mathbf{W}'_{Q_T} \in \mathbb{R}^{d_T \times d_k}$, $\mathbf{W}'_{K_T} \in \mathbb{R}^{d_T \times d_k}$ and $\mathbf{W}'_{V_T} \in \mathbb{R}^{d_T \times d_v}$.

Compared to single head attention, previous work has shown that multi-head attention can make more efficient use of context information Vaswani et al. (2017). The $\mathbf{Q}'_T$, $\mathbf{K}'_T$, and $\mathbf{V}'_T$ are divided into H different subspaces by exploiting H different, learnable linear projections. The Queries, Keys, and Values of the $h-th$ head can be represented as $\mathbf{Q}'_{T,h} = \mathbf{Q}'_T \times \mathbf{W}'_{Q_T,h}$, $\mathbf{K}'_{T,h} = \mathbf{K}'_T \times \mathbf{W}'_{K_T,h}$ and $\mathbf{V}'_{T,h} = \mathbf{V}'_T \times \mathbf{W}'_{V_T,h}$, respectively. Note that $\mathbf{W}'_{Q_T,h} \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$, $\mathbf{W}'_{K_T,h} \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$, $\mathbf{W}'_{V_T,h} \in \mathbb{R}^{d_T \times \frac{d_v}{H}}$. The unimodal self-attention function of $h-th$ head $Att_{T \rightarrow T,h} \in \mathbb{R}^{L_T \times \frac{d_v}{H}}$ is defined as follows.

$$Att_{T \rightarrow T,h} = \mathrm{softmax}\left[\mathbf{Q}'_{T,h} \times \left(\mathbf{K}'_{T,h}\right)^{\top} \Big/ \sqrt{d_k/H}\right] \times \mathbf{V}'_{T,h} \tag{9}$$

The outputs of all the heads are concatenated together and then are linearly transformed to form multi-head unimodal self-attention function:

$$Att^{mul}_{T \rightarrow T} = \mathrm{Concat}\left[Att_{T,0}, Att_{T,1}, \cdots, Att_{T,H}\right] \times \mathbf{W}_{mul} \tag{10}$$

where $Att^{mul}_{T \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$, $\mathbf{W}_{mul} \in \mathbb{R}^{d_v \times d_v}$.

4.3.3. Crossmodal and unimodal attention residual network

After introducing some preliminary definitions, we will present the structure of CARN module in detail. The target modality

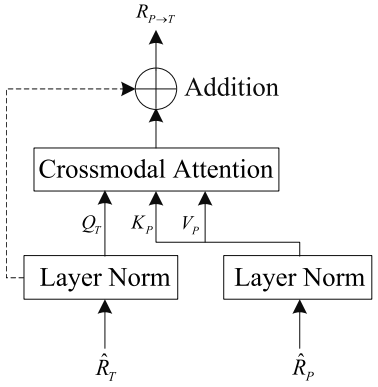

Fig. 2. An illustration for CARN.

selectively extract information from another source modality by exploiting crossmodal attention network. Then, the information is added to the target modality with residual connection. We take pass information from an attached image of the sentence T to the sentence T (i.e., $P \to T$) as an example to make introductions. The overall architecture of the CARN is shown in Fig. 2. To make use of the order of the sequence T, the temporal information is added to the sentence by using PE. It can be computed by:

$$\begin{cases} \widehat{\mathbf{R}}_P = \mathbf{R}_P \\ \widehat{\mathbf{R}}_T = \mathrm{PE}(\mathbf{R}_T) + \mathbf{R}_T \end{cases} \tag{11}$$

Next, the CARN module can be computed by:

$$\mathbf{R}_{P\to T} = Att_{P\to T}\left(\mathrm{LN}\left(\widehat{\mathbf{R}}_P\right), \mathrm{LN}\left(\widehat{\mathbf{R}}_T\right)\right) + \mathrm{LN}\left(\widehat{\mathbf{R}}_T\right) \tag{12}$$

where $\mathbf{R}_{P\to T} \in \mathbb{R}^{L_T \times d_v}$ and LN represents layer normalization Ba, Kiros, and Hinton (2016). Similar to CARN, for a D layers unimodal self-attention residual network (UARN) block, each layer i can be computed by:

$$\begin{aligned} \mathbf{R}^{\mathbf{0}}_{\mathbf{T}\to\mathbf{T}} &= \widehat{\mathbf{R}}_{\mathbf{T}} \\ \mathbf{R}^{i}_{T\to T} &= Att^{mul}_{T\to T}\left(\mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right), \mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right)\right) + \mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right) \end{aligned} \tag{13}$$

where $\mathbf{R}^{i}_{T\to T} \in \mathbb{R}^{L_T \times d_v}$. For simplicity, $\mathbf{R}_{pt} = \mathbf{R}_{P\to T}$, $\mathbf{R}_{tt} = \mathbf{R}^{D}_{T\to T}$. When the target modality is visual information, $\mathbf{R}_{tp} = \mathbf{R}_{T\to P}$.

4.4. Feature extractor

4.4.1. Textual feature extractor

We employed a multi-channel and word-word-aligned CNN-based architecture network (i.e., MCN) to extract the key features from textual information (i.e., $\mathbf{R}_{pt}$ and $\mathbf{R}_{tt}$) processed by CARN and UARN module Kim (2014); Wang et al. (2018a). We first align and stack the embedding matrices $\mathbf{R}_{pt}$ and $\mathbf{R}_{tt}$ as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{pt} \\ \mathbf{R}_{tt} \end{bmatrix} \tag{14}$$

where $\mathbf{R} \in \mathbb{R}^{2\times L_T \times d_v}$. The convolutional filters $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ with various windows size l are used to extract information from embedding matrix $\mathbf{R}$ and produce $L_T - l + 1$ new features. When a filter start with $i - th$ word, the new feature can be denoted as:

$$r_i = \sigma(\mathbf{W}_c \cdot \mathbf{R}_{i:i+l-1} + b_W) \tag{15}$$

where σ is Leaky ReLU activation function and b_W is bias term. The same convolutional operation is performed on each possible window of words in this sentence which generates a feature vector.

$$\mathbf{r} = [r_1, r_2, \cdots, r_{L_T-l+1}] \tag{16}$$

Next, we extract the maximum $\tilde{\mathbf{r}} = \max(\mathbf{r})$ by performing the max-over-time pooling operation on the feature vector $\mathbf{r} \in \mathbb{R}^{L_T-l+1}$. Suppose there being n_W different filters with window size l for the sentence, its feature representation $\tilde{\mathbf{r}}^l \in \mathbb{R}^{n_W}$ can be denoted as:

$$\tilde{\mathbf{r}}^l = \left[\tilde{\mathbf{r}}_1, \tilde{\mathbf{r}}_2 \cdots, \tilde{\mathbf{r}}_{n_W}\right] \tag{17}$$

Table 1
The statistics of datasets

Dataset	# of fake news	# of real news	# of images
Tweet	7,898	6,026	514
Weibo A	4,103	3,605	7,708
Weibo B	5,076	5,008	10,084
Weibo C	5,065	5,065	10,130

Suppose there being n_l filters with different size (i.e., $l \in [1, \cdots, n_l]$),

$$\tilde{\mathbf{R}} = \text{Concat}\left[\tilde{\mathbf{r}}^1, \cdots, \tilde{\mathbf{r}}^{n_l}\right] \tag{18}$$

where $\tilde{\mathbf{R}} \in \mathbb{R}^{d_R}$ and $d_R = n_W \times n_l$. For the sentence T, the final textual feature representation $\mathbf{R}_{tf} \in \mathbb{R}^{d_T}$ is:

$$\mathbf{R}_{tf} = \sigma\left(\mathbf{W}_T \times \tilde{\mathbf{R}} + b_T\right) \tag{19}$$

where σ is Leaky ReLU activation function, $\mathbf{W}_T$ is weight matrix and b_T is bias term.

4.4.2. Visual feature extractor

As a post only with an attached image in our model, we adopted the output (i.e., $\mathbf{R}_{ip}$) of CARN module as the final visual feature representation $\mathbf{R}_{pf} \in \mathbb{R}^{d_P}$.

4.5. Fake news predictor and model learning

We have introduced the mainly modules of this paper. $\mathbf{R}_{tf}$ and $\mathbf{R}_{pf}$ are concatenated together and then are fed to a softmax layer to make the final prediction. The fake news predictor is defined as:

$$\hat{y} = \text{softmax}\left(\mathbf{W} \times \left[\mathbf{R}_{tf}, \mathbf{R}_{pf}\right] + b\right) \tag{20}$$

where $\mathbf{W}$ is the parameters of softmax layer, b is bias term and $\hat{y} = [\hat{y}_0, \hat{y}_1]$. $\hat{y}_0$ and $\hat{y}_1$ denote the probability of a given news is real(0) or fake(1), respectively. We adopt cross-entropy to define the loss function $L(\theta)$ as follows.

$$L(\theta) = -y\log\left(\hat{y}_1\right) - (1-y)\log\left(\hat{y}_0\right) \tag{21}$$

where θ is model parameters. The model aim at minimizing the loss function L_θ for each news by learning θ through back-propagation. We use stochastic gradient decent to train the model and choose Adam as the optimizer with learning rate decay.

5. Experiments

In this section, first, we present the information of four large social media datasets. Second, we provide an introduction about model settings and baseline methods. Third, we make comparisons between the model and baseline methods on four datasets and then bring an detail analysis for experimental results.

5.1. Datasets

The Twitter and Weibo A multimodal dataset are widely adopted by previous work. In addition, in this work, we introduce two new multi-modal fake news datasets for the first time. The detailed statistics information of four datasets is shown in Table 1.

- Twitter Dataset. The Twitter dataset derives from the Verifying Multimedia Use task, the goal of which is to distinguish fake/real news on Twitter with automatic method Boididou et al. (2016). It consists of development set and test set. There is no overlapping events between development set and test set. For each piece of data, it contains text component, an associated image/video and additional user profile information. We only keep the data with text content and attached images.
- Weibo A Dataset. The Weibo A datatset is first presented in Jin et al. (2017a) for fake news detection task. Each post contains text content, user profiles and attached images. The verified fake news is collected from the official fake news debunking system of the Sina Weibo, a website very similar to Twitter. The time span of the data is from May 2012 to January 2016. Jin et al. adopted the news verified by the Xinhua News Agency as real news Jin et al. (2017a). Following the data preprocessing methods of previous research, the low quality and duplicated images are taken away Slaney and Casey (2008). To avoid the same events among training, validation and testing set, we find the same events by exploiting a single-pass clustering method Yang, Pierce, and Carbonell (1998).

- Weibo B Dataset. The Weibo B dataset, a benchmark dataset for internet fake news detection challenger[3], is released by Cao et al. Cao et al. (2018). For each post, it contains text content, attached images, user profile information, news category and corresponding ground-truth label. We take the way same to Weibo A dataset to preprocess and split the Weibo B dataset.
- Weibo C Dataset. To promote fake news detection task, we build a new multi-modal fake news detection dataset. The fake news is collected from the Weibo community management center[4], an official fake news debunking system. The time span of this data is from May 2012 to November 2019. The real news is collected from the People's Daily, an authoritative news source similar to the Xinhua News Agency. Each post contains the original post text, associated images/video and additional user profile information. Following previous research, we preprocess and split this dataset with the way same to Weibo A and B datasets.

5.2. Experimental setup

Following previous work Wang et al. (2018b), development set and test set of the Tweet dataset is used as training set and test set, respectively. For each Weibo datasets, we choose 70%, 10% and 20% of news for training, validation and testing set, respectively. Same to previous research, we obtain 32-dimensional word embedding (i.e., $d_T = 32$) by exploiting Word2Vec model Mikolov, Sutskever, Chen, Corrado, and Dean (2013). For textual feature extractor, the window size of the filter is $l \in [1, 2, 3, 4]$ (i.e., $n_l = 4$) and for each size l, the number of filters is $n_W = 25$. For CARN, the number of layers and heads has little effect on the experimental results, so we choose single head and layer attention residual network. For UARN, we choose the attention residual network with 4 heads and 3 layers (i.e., $H = 4$ and $D = 3$), which achieves the best performance. In the process of training, the batch size and the number of epochs is set to 150. We choose Accuracy, Precision, Recall and F_1 score as evaluation metrics which are widely adopted by related areas Shu et al. (2017).

5.3. Baselines

We make comparisons with a series of baseline fake news detection methods as follows.

5.3.1. Single modality models

- Textual. As the input of model is only post, the CARN module is removed. The output of sentence embedding layer is fed into single channel CNN-based textual feature extractor Kim (2014), which followed by a fully connected layer and softmax layer.
- Visual. The visual features are obtained from the VGG-19 net. After processed by input embedding layer, The visual information is fed into a fully connected layer and softmax layer for making final prediction.

5.3.2. Multimodal models

- VQA Antol et al. (2015). The goal of Visual Question Answering (VQA) is to provide an answer to a question about a given image. As VQA is a multi-classification task, we have to replace the multi-class classifier with a binary classifier. For a fair comparison, we choose a single layer LSTM with hidden layer size 32.
- NeuralTalk Vinyals, Toshev, Bengio, and Erhan (2015). The NeuralTalk model is proposed to generate natural language descriptions from visual information. To adapt the model to fake news detection task, its feature representation is defined as the average of the output of RNN at each time step. Then, the feature representations are fed into a fully connected layer to make prediction. For a fair comparison, we choose both LSTM and fully connected with the hidden layer size 32.
- att-RNN Jin et al. (2017a). att-RNN is a RNN-based automatic multimodal fake news detection model which fuses joint representation of textual features and user profile features and visual features via attention mechanism. For a fair comparison, the user profile information is removed and the hidden layer size of LSTM is 32.
- EANN Wang et al. (2018b). The Event Adversarial Neural Networks (EANN) is a multi-task learning fake news detection model, which aims at learning shared feature representations among all the posts by leveraging an additional adversarial component. Textual and visual feature representations are obtain by exploiting a CNN-based textual features extractor Kim (2014) and VGG-19 network, respectively. For a fair comparison, we remove the adversarial component.
- MVAE Khattar et al. (2019). The state-of-the-art method, the Multimodal Variational Autoencoder (MVAE), is a multi-task learning multimodal fusion fake news detection framework. The modal aims at discovering correlations across modalities by exploiting VAE to reconstructs the textual and visual feature representations from the shared latent representation.
- MKN Zhang et al. (2019). Multi-modal Knowledge-aware Event Memory Network (MKEMN) is event-level multi-modal fake news detection framework, which use the visual information and the external knowledge to assist fake news detection task. The authors adopted an event memory network to learn event invariant features. Considering the differences between event-level and post-level fake news detection and the fairness of comparison, we remove the external knowledge component and event memory network. The modified method is denoted as MKN.

[3] https://biendata.com/competition/falsenews/
[4] http://service.account.weibo.com/

Table 2
The experimental results of different methods on Twitter dataset.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Textual	0.568	0.655	0.379	0.480	0.531	0.778	0.631
	Visual	0.664	0.733	0.568	0.640	0.617	0.770	0.685
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.650
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.681	0.769	0.561	0.650	0.626	0.813	0.707
	EANN	0.677	0.750	0.579	0.653	0.627	0.786	0.699
	MVAE	0.578	0.626	0.488	0.548	0.544	0.677	0.603
	MKN	0.664	0.753	0.537	0.627	0.611	0.805	0.695
	CARMN	**0.741**	**0.854**	**0.619**	**0.718**	**0.670**	**0.880**	**0.760**

Table 3
The experimental results of different methods on three Weibo datasets.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo A	Textual	0.764	0.776	0.721	0.747	0.755	0.805	0.779
	Visual	0.594	0.583	0.752	0.657	0.615	0.424	0.502
	VQA	0.579	0.581	0.665	0.620	0.576	0.487	0.527
	Neural Talk	0.748	0.739	0.790	0.764	0.758	0.702	0.730
	att-RNN	0.784	0.797	0.781	0.789	0.771	0.787	0.779
	EANN	0.807	0.831	0.788	0.809	0.785	0.828	0.806
	MVAE	0.681	0.756	0.589	0.662	0.630	0.785	0.698
	MKN	0.792	0.805	0.788	0.796	0.778	0.796	0.787
	CARMN	**0.853**	**0.891**	**0.814**	**0.851**	**0.818**	**0.894**	**0.854**
Weibo B	Textual	0.762	0.861	0.623	0.723	0.706	0.900	0.791
	Visual	0.702	0.734	0.630	0.678	0.678	0.773	0.722
	VQA	0.704	0.706	0.695	0.701	0.702	0.713	0.707
	Neural Talk	0.735	0.778	0.652	0.709	0.704	0.817	0.756
	att-RNN	0.780	0.853	0.675	0.753	0.733	0.884	0.801
	EANN	0.815	0.903	0.703	0.791	0.759	0.925	0.834
	MVAE	0.741	0.779	0.671	0.721	0.713	0.811	0.759
	MKN	0.778	0.880	0.643	0.743	0.720	0.913	0.805
	CARMN	**0.869**	**0.935**	**0.796**	**0.860**	**0.820**	**0.944**	**0.878**
Weibo C	Textual	0.772	0.742	0.844	0.790	0.812	0.697	0.750
	Visual	0.831	0.806	0.882	0.842	0.864	0.779	0.820
	VQA	0.807	0.742	0.953	0.834	0.931	0.657	0.770
	Neural Talk	0.796	0.751	0.897	0.817	0.867	0.691	0.769
	att-RNN	0.834	0.778	0.942	0.852	0.923	0.722	0.810
	EANN	0.858	0.807	0.948	0.872	0.934	0.765	0.841
	MVAE	0.821	0.781	0.901	0.837	0.878	0.737	0.802
	MKN	0.842	0.786	0.947	0.859	0.930	0.733	0.820
	CARMN	**0.922**	**0.890**	**0.965**	**0.926**	**0.961**	**0.876**	**0.917**

5.4. Results and analysis

5.4.1. Comparisons of different models

Table 2 and Table 3 show the performance of the proposed model as well as baseline methods in fake news detection task on Twitter and Weibo dataset, respectively. We can observe that CARMN outperforms all the competitive models on different metrics. In fact, the Tweet dataset is not a good choice for the post-level fake news detection task. There are multiple languages, many irregularly written texts, and the textual features lack diversity, which is the reason that the performance of the method based on textual information is worse than visual information. By fusing the information across modalities via attention mechanism, att-RNN outperforms all baseline methods, which confirms the effectiveness of using multimodal information simultaneously in the fake news detection task. Compared with the Tweet dataset, there is rich semantic context information in the text of the three Chinese datasets.

On Chinese datasets, the method only based on text features shows similar or even better performance than some baseline methods. There is a phenomenon that can not be ignored is that, on Weibo C dataset, the method based on visual features outperforms the method based on textual feature representation, which can be attributed to the high-level of quality of the pictures attached to the real news. For multimodal models, attention-based methods (i.e., att-RNN and MKN) show better performance than VQA and NeuralTalk but are less effective and robustness than EANN. MVAE shows worse performance, the reason of which is that it only adopts shared latent representation between textual and visual information. It suggests that it is important to keep unique characteristics for each modality. By introducing the CARN and MCN, our model can keep unique characteristics for each modality while fusing the correlative

Table 4
The comparison of experimental results among variants of CARMN.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
CNN*	0.862	0.815	0.943	0.874	0.930	0.777	0.847
CARN*	0.890	0.850	0.953	0.898	0.944	0.825	0.881
CARMN-	0.913	0.878	0.963	0.919	0.959	0.861	0.907
CARMN	**0.922**	**0.890**	**0.965**	**0.926**	**0.961**	**0.876**	**0.917**

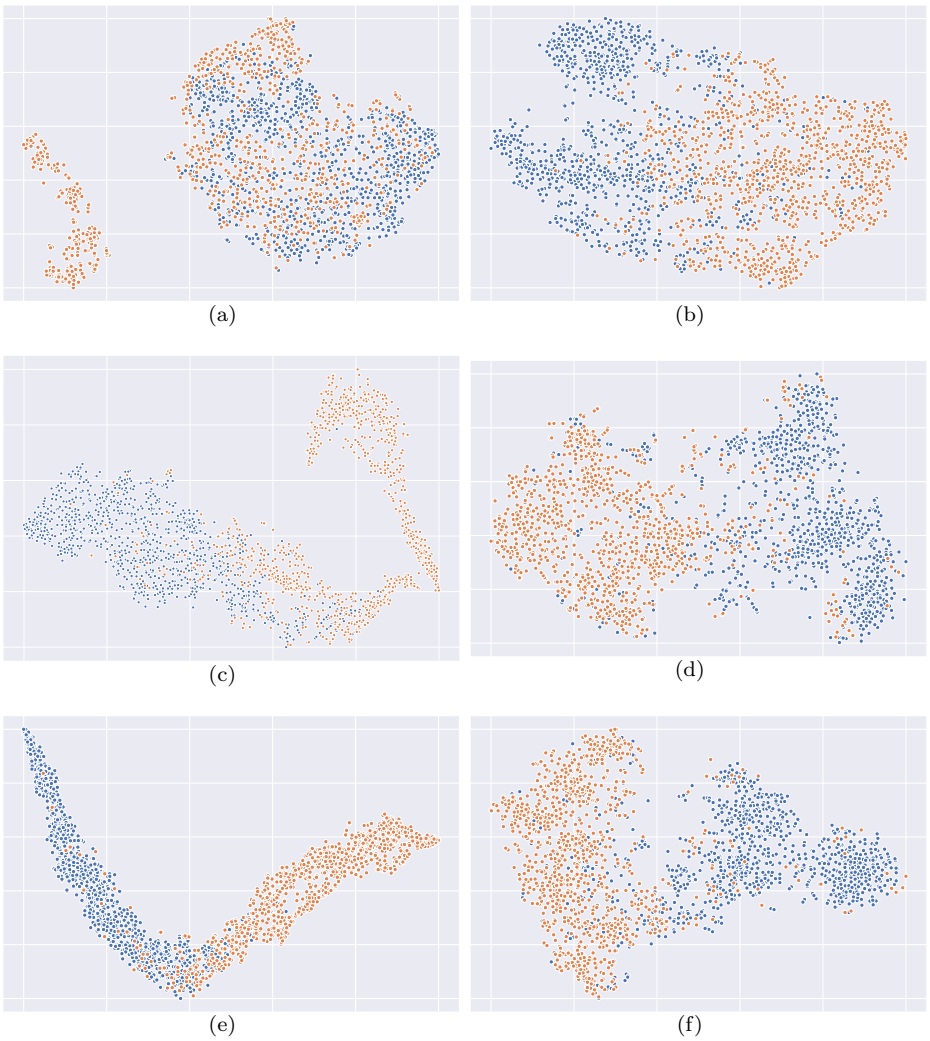

Fig. 3. Visualization of learned latent textual and visual feature representations on the testing data of Weibo C dataset with *t*-SNE, (a) the textual feature representation $\mathbf{R}_{tf}$ learned by CNN*; (b) the visual feature representation $\mathbf{R}_{pf}$ learned by CNN*; (c) the textual feature representation $\mathbf{R}_{tf}$ learned by CARN*; (d) the visual feature representation of $\mathbf{R}_{pf}$ learned by CARN*; (e) the textual feature representation $\mathbf{R}_{tf}$ learned by CARMN; (f) the visual feature representation of $\mathbf{R}_{pf}$ learned by CARMN.

and complementary information between different modalities and alleviate the influence of noise information which may be generated by crossmodal fusion component. The experimental results demonstrate the effectiveness of the proposed model.

5.4.2. Comparisons among variants of CARMN

To further validate the effectiveness of CARMN, we make comparisons with variants of CARMN as follows.

Table 5
The comparison of experimental results between CARMN and SpotFake.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
SpotFake	0.848	0.804	0.914	0.855	0.904	0.784	0.840
CARMN_Bert	**0.934**	**0.922**	0.952	**0.937**	0.948	**0.916**	**0.932**
CARMN	0.922	0.890	**0.965**	0.926	**0.961**	0.876	0.917

Table 6
The comparison of experimental results between CARMN and SpotFake+.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
SpotFake+	0.838	0.807	0.882	0.843	0.875	0.796	0.834
CARMN_XLNet	0.922	0.890	**0.965**	**0.926**	**0.961**	0.876	0.917
CARMN	**0.924**	**0.921**	0.930	0.925	0.926	**0.917**	**0.922**

Table 7
The experimental results of different layer number.

# of layer	1	2	3	4	5
Accuracy	0.915	0.917	0.922	0.914	0.906
F1 Score	0.914	0.917	0.921	0.914	0.906

(1) CNN*: The CNN* is the variant of CARMN. It is removed the CARN and UARN. The convolutional filters is changed from $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ to $\mathbf{W}_c \in \mathbb{R}^{1\times l\times d}$.
(2) CARN*: The CARN* is the variant of CARMN. It is removed the UARN. The convolutional filters is changed from $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ to $\mathbf{W}_c \in \mathbb{R}^{1\times l\times d}$.
(3) CARMN-: The CARMN- is the variant of CARMN. It is removed the unimodal attention module.

Due to space limitations, all of the subsequent experimental analyses only focus on the Weibo C dataset. The results are shown in the Table 4. The CARN* is better that CNN* which proves that by fusing the correlative information between different modalities can benefit the model's performance. Compared with CARN*, the usage of the additional residual network and MCN (i.e., CARMN-) can further improve the accuracy and mitigate the influence of noise information which may be generated by crossmodal fusion component. We can achieve the best performance by combining CARMN- with the self-attention module. Then, as shown in the Fig. 3, we visualize the final feature representation $\mathbf{R}_{tf}$ (i.e., a, c, e) and $\mathbf{R}_{pf}$ (i.e., b, d, f) learned by CNN*, CARN* and CARMN with t-SNE van der Maaten and Hinton (2008). The orange and blue color nodes represent fake and real news, respectively. We can observe that our CARMN learns more discriminable feature representations. For textual feature representation, the rank of its discriminability is Fig. 3 (e) > Fig. 3(c) > Fig. 3(a). For visual feature representation, the rank of its discriminability is Fig. 3(f) > Fig. 3(d) > Fig. 3(b). It proves that the CARMN can learn more discriminability feature representations and further validate the effectiveness of the proposed method. The reason why we not to visualize the CARMN- is that the results of CARMN and CARMN- are similar.

5.4.3. Comparisons with transfer learning-based methods

In addition, we also make comparisons with transfer learning-based methods and launch an investigation to find out how the pre-trained Bert Devlin, Chang, Lee, and Toutanova (2018) and XLNet Yang et al. (2019) affect the proposed model's performance. The SpotFake Singhal et al. (2019) and SpotFake+ Singhal et al. (2020), transfer learning-based fake news detection methods, are mainly based on Bert and XLNet model, respectively. The CARMN that takes word embedding representation from Word2Vec model Mikolov et al. (2013) is replaced by CARMN_Bert and CARMN_XLNet that take the representation from pre-trained Bert and XLNet with no fine-tuning. The experimental results are shown in Table 5 and Table 6. Compared with SpotFake and SpotFake+, CARMN_Bert and CARMN_XLNet show better performance. However, the experimental results of CARMN_Bert and CARMN_XLNet are similar to CARMN, which shows that the word embedding representation from pre-trained Bert and XLNet model fail to largely improve the model's performance.

5.4.4. Effects of the number of the heads and layers

In this section, we investigate how the number of the self-attention heads and residual network layers affect the model's performance. Specifically, we set the range of the number of layer to $[1, 2, 3, 4, 5]$. Table 7 shows the performance of CARMN with different layers. The performance of CARMN increases with the number of the layers grows until 3. As the number of the self-attention head

Table 8
The experimental results of different head number.

# of head	1	2	4	8
Accuracy	0.917	0.919	0.922	0.922
F1 Score	0.917	0.919	0.921	0.922

must be divisible by word embedding dimension, we set the range of the number of the head to $[1, 2, 4, 8]$. Table 8 shows the performance of CARMN with different heads. We can observe that the performance of CARMN increases with the number of heads grows until 4. That's the reason why we set the number of the self-attention heads and residual network layers as 4 and 3, respectively.

6. Conclusion

In the field of multimodal fake news detection, there is a challenge of keeping the unique properties for each modality while fusing the relevant information between different modalities. However, for some posts and their attached images, the fusion between textual and visual information may produce noise information which may affect model's performance. To solve these problems, we proposed a multimodal fake news detection model based on CARN and MCN. We conduct extensive experiments on four real-world datasets and demonstrate the effectiveness of the proposed model. As the CARMN is a general model for multimodal fake news detection task, it can be easily expanded to more modalities and the multimodal fusion module can be replaced by other methods. In future work, we will explore event-level multimodal fake news detection by exploiting visual information.

CRediT authorship contribution statement

Chenguang Song: Conceptualization, Methodology, Software, Validation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Nianwen Ning:** Resources, Data curation, Software, Writing - review & editing. **Yunlei Zhang:** Writing - original draft, Writing - review & editing. **Bin Wu:** Supervision.

Acknowlgedgments

This work is supported by the National Key Research and Development Program of China (Grant No. 2018YFC0831500), National Natural Science Foundation of China (Grant No. 61972047) and the NSFC-General Technology Basic Research Joint Funds (Grant No. U1936220).

References

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on twitter with hybrid cnn and rnn models, . In *SMSociety '18Proceedings of the 9th international conference on social media and society* (pp. 226–230). ACM. https://doi.org/10.1145/3217804.3217917.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: visual question answering. *The ieee international conference on computer vision* (pp. 2425–2433). https://doi.org/10.1007/s11263-016-0966-6.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization.

Boididou, C., Papadopoulos, S., Dang-Nguyen, D., Boato, G., Riegler, M., Middleton, S. E., … Kompatsiaris, Y. (2016). Verifying multimedia use at mediaeval 2016. *Working notes proceedings of the mediaeval 2016 workshop*. CEUR-WS.org.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences, 497*, 38–55. https://doi.org/10.1016/j.ins.2019.05.035.

Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., & Li, J. (2018). Automatic rumor detection on microblogs: a survey.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., & Li, J. (2020). *Exploring the role of visual content in fake news detection*. In K. Shu, S. Wang, D. Lee, & H. Liu (Eds.) (pp. 141–161)). Cham: Springer International Publishing.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th international conference on world wide web* (pp. 675–684). ACM. https://doi.org/10.1145/1963405.1963500.

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: deep attention based recurrent neural networks for early rumor detection. *Proceedings of the 22nd pacific-asia conference on knowledge discovery and data mining* (pp. 40–52). Springer International Publishing. https://doi.org/10.1007/978-3-030-04503-6_4.

Chen, Y., Sui, J., Hu, L., & Gong, W. (2019). Attention-residual network with cnn for rumor detection. *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 1121–1130). ACM. https://doi.org/10.1145/3357384.3357950.

Cui, L., Wang, S., & Lee, D. (2019). Same: Sentiment-aware multi-modal embedding for detecting fake news. *Proceedings of the 2019 ieee/acm international conference on advances in social networks analysis and mining* (p. 4148). Association for Computing Machinery. https://doi.org/10.1145/3341161.3342894.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dungs, S., Aker, A., Fuhr, N., & Bontcheva, K. (2018). Can rumour stance alone predict veracity?. *Proceedings of the 27th international conference on computational linguistics* (pp. 3360–3370). Association for Computational Linguistics.

Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 943–951). ACM. https://doi.org/10.1145/3269206.3271709.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. *Proceedings of the 22nd international conference on world wide web* (pp. 729–736). ACM. https://doi.org/10.1145/2487788.2488033.

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on twitter. *Proceedings of the 2012 siam international conference on data mining* (pp. 153–164). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972825.14.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017a). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *Proceedings of the 25th acm international conference on multimedia* (pp. 795–816). ACM. https://doi.org/10.1145/3123266.3123454.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. *Proceedings of the thirtieth aaai conference on artificial intelligence* (pp. 2972–2978). AAAI Press.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017b). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia, 19*(3), 598–608. https://doi.org/10.1109/TMM.2016.2617078.

Kakol, M., Nielek, R., & Wierzbicki, A. (2017). Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management, 53*(5), 1043–1061. https://doi.org/10.1016/j.ipm.2017.04.003.

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. *Proceedings of the 27th international conference on computational linguistics* (pp. 1546–1557). Association for Computational Linguistics.

Ke, W., Song, Y., & Kenny Q, Z. (2015). False rumors detection on sina weibo by propagation structures. *Ieee 31st international conference on data engineering* (pp. 651–662). IEEE. https://doi.org/10.1109/ICDE.2015.7113322.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: multimodal variational autoencoder for fake news detection. *Proceedings of the 28th international conference on world wide web* (pp. 2915–2921). ACM. https://doi.org/10.1145/3308558.3313552.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: multi-task learning for rumour verification. *Proceedings of the 27th international conference on computational linguistics* (pp. 3402–3413). Association for Computational Linguistics.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 3818–3824). AAAI Press.

Ma, J., Gao, W., & Wong, K.-F. (2018a). Detect rumor and stance jointly by neural multi-task learning. *Companion proceedings of the the web conference 2018* (pp. 585–593). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3188729.

Ma, J., Gao, W., & Wong, K.-F. (2018b). Rumor detection on twitter with tree-structured recursive neural networks. *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 1980–1989). Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1184.

Ma, J., Gao, W., & Wong, K.-F. (2019). Detect rumors on twitter by promoting information campaigns with generative adversarial learning. *Proceedings of the 28th international conference on world wide web* (pp. 3049–3055). ACM. https://doi.org/10.1145/3308558.3313741.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th international conference on neural information processing systems* (pp. 3111–3119). Curran Associates Inc.. https://doi.org/10.5555/2999792.2999959.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection.

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. *2019 ieee international conference on data mining*. IEEE.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 3834–3840). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/533.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (pp. 83:1–83:4). American Society for Information Science.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 797–806). ACM. https://doi.org/10.1145/3132847.3132877.

Sejeong, K., Meeyoung, C., Kyomin, J., Wei, C., & Yajun, W. (2013). Prominent features of rumor propagation in online social media. *Ieee 13th international conference on data mining* (pp. 1103–1108). IEEE. https://doi.org/10.1109/ICDM.2013.61.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: a survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology, 10*(3), 21:1–21:42. https://doi.org/10.1145/3305260.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). defend: explainable fake news detection. *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 395–405). ACM. https://doi.org/10.1145/3292500.3330935.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 2236. https://doi.org/10.1145/3137597.3137600.

Shu, K., Wang, S., & Liu, H. (2018). Understanding user profiles on social media for fake news detection. *Proceedings - ieee 1st conference on multimedia information processing and retrieval* (pp. 430–435). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/MIPR.2018.00092.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International conference on learning representations*.

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Aaai* (pp. 13915–13916).

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. *2019 ieee fifth international conference on multimedia big data (bigmm)* (pp. 39–47).

Slaney, M., & Casey, M. (2008). Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal processing magazine, 25*(2), 128–131.

Su, T., Macdonald, C., & Ounis, I. (2019). Ensembles of recurrent networks for classifying the relationship of fake news titles, . In *SIGIR19Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (p. 893896). ACM. https://doi.org/10.1145/3331184.3331305.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6558–6569). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1656.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc..

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: a neural image caption generator. *The ieee conference on computer vision and pattern recognition* (pp. 3156–3164). https://doi.org/10.1109/CVPR.2015.7298935.

Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. *The 41st international acm sigir conference on research & development in information retrieval* (pp. 275–284). ACM. https://doi.org/10.1145/3209978.3210037.

Vo, N., & Lee, K. (2019). Learning from fact-checkers: analysis and generation of fact-checking language. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 335–344). ACM. https://doi.org/10.1145/3331184.3331248.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559.

Wang, H., Zhang, F., Xie, X., & Guo, M. (2018a). Dkn: Deep knowledge-aware network for news recommendation. *Proceedings of the 2018 world wide web conference* (pp. 1835–1844). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186175.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., … Gao, J. (2018b). Eann: event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). ACM. https://doi.org/10.1145/3219819.3219903.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the acm sigkdd workshop on mining data semantics* (pp. 13:1–13:7). ACM. https://doi.org/10.1145/2350190.2350203.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study of retrospective and on-line event detection, . In *SIGIR 98Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 28–36). ACM. https://doi.org/10.1145/290941.290953.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 5753–5763). Curran Associates, Inc.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 3901–3907). AAAI Press. https://doi.org/10.24963/ijcai.2017/545.

Zhang, H., Fang, Q., Qian, S., & Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. *Proceedings of the 27th acm international conference on multimedia* (pp. 1942–1951). ACM. https://doi.org/10.1145/3343031.3350850.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management, 57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., & Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management, 56*(6), 102097. https://doi.org/10.1016/j.ipm.2019.102097.

Zhou, K., Shu, C., Li, B., & Lau, J. H. (2019a). Early rumour detection. *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1614–1623). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1163.

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019b). Fake news: fundamental theories, detection strategies and challenges. *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 836–837). ACM. https://doi.org/10.1145/3289600.3291382.