Journal of New Media
DOI: 10.32604/jnm.2022.027312
Article

Tech Science Press

Cross-Modal Relation-Aware Networks for Fake News Detection

Hui Yu and Jinguang Wang*

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Jinguang Wang. Email: wangjinguang502@gmail.com
Received: 14 January 2022; Accepted: 09 March 2022

Abstract: With the speedy development of communication Internet and the widespread use of social multimedia, so many creators have published posts on social multimedia platforms that fake news detection has already been a challenging task. Although some works use deep learning methods to capture visual and textual information of posts, most existing methods cannot explicitly model the binary relations among image regions or text tokens to mine the global relation information in a modality deeply such as image or text. Moreover, they cannot fully exploit the supplementary cross-modal information, including image and text relations, to supplement and enrich each modality. In order to address these problems, in this paper, we propose an innovative end-to-end Cross-modal Relation-aware Networks (CRAN), which exploits jointly models the visual and textual information with their corresponding relations in a unified framework. (1) To capture the global structural relations in a modality, we design a global relation-aware network to explicitly model the relation-aware semantics of the fragment features in the target modality from a global scope perspective. (2) To effectively fuse cross-modal information, we propose a cross-modal co-attention network module for multi-modal information fusion, which utilizes the intra-modality relationships and inter-modality relationship jointly among image regions and textual words to replenish and heighten each other. Extensive experiments on two public real-world datasets demonstrate the superior performance of CRAN compared with other state-of-the-art baseline algorithms.

Keywords: Fake news detection; relation-aware networks; multi-modal fusion

1 Introduction

Recently, with the great development of the communication Internet, social multimedia has been increasingly extensive in our life. Due to the straightforward accessibility, people often select social multimedia to obtain and express as well as change views. Unfortunately, owing to the wide scale of users and the many complex sources, it is easy to generate different kinds of fake news. With wide-ranging of fake news being exploited incorrectly for deceiving users, in some cases may bring about severe real-world threats on society and cause serious economic losses. Most users don't have time and

energy to check the credibility of each piece of information on the Internet. Hence, it is imperative to identify the fake news from social multimedia in time and guarantee users receive truthful information.

To date, there are different kinds of fake news detection methods [1–4] being researched, which contains not only traditional machine learning but also deep learning-based approaches. The former such as Random Forest, Decision Tree and Support Vector Machine (SVM) depend on hand-craft representations to spot fake news, leading to labor surplus and time-consuming. For instance, SVM-TS [1] employs heuristic rules along with the linear SVM [5] classifier to predict rumors real or not on Twitter. Also, it utilizes time-series structural information to capture the variations of social representations. With the huge success of the neural networks, most existing deep learning methods have already achieved better improvement than the traditional ones in terms of model performance owing to the advanced capability of feature extraction. Some early works attempted to learn feature representations from pure text content to perform detection. Further, it sought to use Recurrent Neural Networks (RNN) [4] along with its variants. For instance, Long Short-Term Memory (LSTM) unit is employed to learn temporal semantic representations for debunking fake news. Based on these works, another work attempts to introduce the attention mechanism into RNNs to learn temporal semantic feature representations of concrete focal points for detection. Besides, some works also introduce Convolutional Neural Networks (CNN) [2] to capture the high-level features learned from posts to discover fake news. However, these methods based on single-modality have achieved some successes, they can only explore the local feature relation representations in intra-modality such as textual modality via conducting a weighted summation of the feature representations from all positions to the given target position simply, and they usually ignore the global relation information between all the feature segments and the target ones of posts, which can learn better semantics to identify fake news.

Nowadays, the content of news has evolved from pure text to multi-modal content containing texts, images and so on due to the development of social multimedia. Fig. 1 illustrates a case of the multi-modal news. Multi-modal fake news detection has already gained concerns increasingly. Many studies [6–9] adopt deep learning modules to capture and integrate visual and textual feature representations of posts. Some models [7,8] only simply concatenate representations extracted from image and text modality jointly to generate the final features. Others take inter-modality attention to complement and enhance semantic concepts in one modality space. However, these methods could not make full use of visual information, which can enrich the semantics between different modalities. Also, these fundamental approaches cannot effectively combine supplementary and relational multi-modal information among image and text fragments to enrich and heighten each other.

It seems that the existing methods have achieved better performance. But they also face some challenges, which can be described as follows:

- challenge 1: How to explicitly model the global spatial position binary relations among fragments of the posts in intra-modality to obtain a valuable global scope structural relational information for debunking fake news?
- challenge 2: How to integrate complementary cross-modal information including semantic concepts and entities between different modalities to enhance the efficiency of detecting fake news?

Figure 1: A case of the multi-modal news on social multimedia

To deal with these above challenges, we propose an end-to-end Cross-modal Relation-aware Networks (CRAN) to use the global relation-aware module to capture the global spatial connection features among fragments in each modality, and then utilizing cross-modal co-attention module to incorporate complementary and noisy multi-modal information with global relation-aware module jointly for fake news detection. In order to gain a robust model, we design two adjacent modules containing global relation-aware network and cross-modal co-attention network: (1) For challenge 1, we employ a global relation-aware network to explicitly learn global spatial scope relations to mine the global structural relational features in each modality. (2) For challenge 2, to effectively mine supplementary and relational multi-modality information consisting of image and text segments semantics, we design a cross-modal co-attention network on the basis of captured fine-grained features of image and text segments with their global relational semantics. By jointly taking inter-modality and intra-modality relations jointly, we can supplement and enrich each modality of posts in a common semantic space.

To sum up, the contributions of our work are as follows:

- We propose an innovative end-to-end Cross-modal Relation-aware Networks (CRAN) to jointly model the global relation scope information of each intra-modality and multi-modal information consisting of image and text fragments into a unified model for performing fake news detection tasks.
- A global relation-aware network is used to obtain the global spatial scope binary relations among the fragment feature nodes of each modality of the social multimedia posts content, which can be captured via pre-trained models, and apply two convolutional layers to derive the attention of each feature node on the basis of the obtained global spatial scope relations. Then, we employ a cross-modal co-attention module for multi-modal fusion, which exploits the intra-modality inter-modality relationship jointly among image and text segments with their global scope relations in each modality between image regions and textual words to supplement and heighten each other for obtaining outstanding comprehensive cross-modal representations of posts.

- We evaluate our model (CRAN) on the two real-world experimental datasets (e.g., PHEME and WEIBO), and the experimental results demonstrate CRAN outperforms the state-of-the-art baseline models.

2 Problem Statement

Generally speaking, the fake news detection task can be defined as a binary classification issue, which aims to predict a post in social media as true news or fake news. Given a group of multi-modal posts $P = \{p_1, p_2, \ldots, p_m\}$, where p_i is a post containing a set of words with its attached image information, m represents the number of given posts. Our goal is to learn a model $F{:}P \rightarrow Y$, to classify every post p_i into the predefined classes $Y = \{0, 1\}$ where 1 denotes real news while 0 denotes fake news.

3 The Proposed Algorithm

3.1 Overall Framework

The goal of our model is to predict whether a post consisting of textual and visual information from social multimedia is real news or fake news. To this end, we propose a Cross-modal Relation-aware Networks (CRAN), which jointly models the textual and visual information along with their global scope relations in a unified model. Fig. 2 displays the framework of CRAN, which mainly contains the following modules:

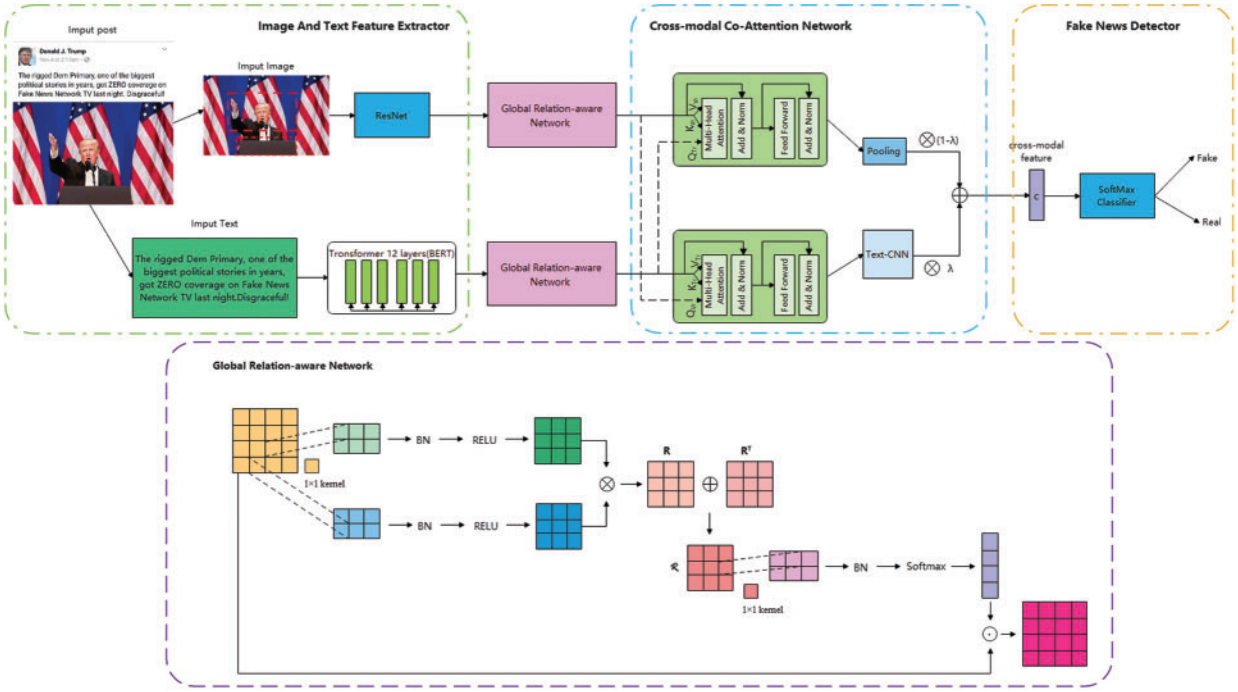

Figure 2: The overall framework of the proposed CRAN

- Image And Text Feature Extractor: Given a multi-modal post on social multimedia including textual and visual information (here refers specifically to the image). For the image branch, we also use a pre-trained model that calls ResNet50 [11] to capture the region features of the image. Simultaneously, for the text branch, we exploit Word Piece embeddings of the sentence words as the segments of the text modality. And then, we utilize another pre-trained model Bidirectional Encoder Representations from Transformers (BERT) [10], to gain the fragment features of the sentence words.
- Global Relation-aware Network: On the basis of the above-obtained fragment features, which are extracted by pre-trained models, we apply a relation-aware attention network to learn a global spatial scope relation feature vector to mine deep global semantic features of each modality of posts regardless of visual modality or textual modality. Specifically, we adopt a 1×1 convolutional layer, followed by which are the Batch Normalization (BN) and ReLU activation function, to capture the global spatial relations for mining the implicit relations in each modality. And further, we will utilize a 1×1 convolutional layer, followed by which are the BN and Softmax activation function, to gain the feature vector weight and then make it multiply with the fragment feature to the global relation features.
- Cross-modal Co-attention Network: On the basis of the captured fine-grained relation features for image and text fragments, we apply a cross-modal co-attention network unit to jointly model the intra-modality and inter-modality relations among image and text segments. By jointly premeditating inter-modality and intra-modality relation information of different modalities, we can improve the ability to extract the feature representations of image and text fragments. Then, we apply the Text-CNN unit and pooling operation to aggregate the learned fine-grained relation representations as the final cross-modal feature representations of posts.
- Fake news Detector: We designed this module is aimed to identify the post collected from social media is fake or not. Then the detector uses the above cross-modal features as the input and applies a fully-connected layer along with its relevant activation function to predict classification probability for detecting fake news.

3.2 Image and Text Feature Extractor

As mentioned above, the input of this module is the multi-modal post $s = \{P, W\}$, where P and W represent the visual and textual content, respectively.

For the visual branch, given a visual image content P, we utilize the pre-trained ResNet50 [11] unit to capture region representations of images $V = \{v_1, v_2, \ldots, v_k\}$, where k represents the number of regions of a given image. Also, each v_i is viewed as the mean-pooling convolutional feature of the i-th region. That is to say, given the affiliated image P, using the penultimate pooling layer of the ResNet50 to extract its region representations, which can be defined as:

$$V = \{v_1, \ldots, v_k\} = ResNet50(P) \tag{1}$$

where $v_i \in \mathbb{R}^{d_p}$ and d_p denotes the dimension of each region representation.

For the textual branch, to accurately extract the semantics of the text, we apply BERT module as the core unit of our text feature extraction, which has been shown to be valid in multiple areas such as translation, reading comprehension and text classification [12–14]. If given a text W of the post, we view W as the sequential list of text words $W = \{w_1, w_2, \ldots, w_n\}$, where n denotes the number of the words in text. Then, we feed it into BERT to gain the transformed representations as $T = \{t_1, t_2, \ldots, t_n\}$, where t_i is viewed as the relevant transformed representation of w_i. The words feature representation

t_i can be learned via BERT as follows:

$$T = \{t_1, \ldots, t_n\} = BERT(W) \tag{2}$$

where $t_i \in \mathbb{R}^{d_w}$ is the corresponding token of the last layer output of BERT unit. And the d_w represents the dimension of the word fragment representation.

During the training stage, these pre-trained models are fixed. Also, they are chosen in parallelism to earlier researches on this issue for comparing the effectiveness of our structure.

It is evident that $d_p \neq d_w$, in order to adapt to our task, the dimension of image fine-grained representations should be the same as that of text word fine-grained representations, represented as $d_p = d_w = d_s$.

3.3 Global Relation-Aware Network

To effectively explore the deep semantic information in each modality, such as the visual modality or textual modality of posts. We design the global relation-aware network (GRN) to realize it. The input of GRN is the fragment features, and the output of GRN is the deep fine-grained features containing global scope relation information.

Specially, for the given fragment features of a modality that can be captured by the pre-trained models, which denoted as $F = \{f_1, f_2, \ldots, f_n\} \in \mathbb{R}^{n \times d_f}$, we employ the relation-aware attention network to generate the deep fine-grained features that include potential global relationship information, which can be shown in Fig. 2. At first, for the lower complexity, we utilize a 1×1 convolutional layer followed by BN and ReLU activation operation one after another to obtain two feature maps. The formulations can be denoted as follows:

$$\begin{aligned} A &= \phi(f_i;\ \theta_\phi) = \mathrm{Re}LU(W_{\theta_\phi} f_i) \\ B &= \phi(f_i;\ \theta_\phi) = \mathrm{Re}LU(W_{\theta_\phi} f_i) \end{aligned} \tag{3}$$

Note that in order to simplify the notation, the BN operation is omitted. These two characteristic matrices that denoted as A and B, conform to the standard normal distribution. Then, we apply the multiplication between the two matrices to get the relation matrix:

$$R = A^T B = (r_{i,j}) \tag{4}$$

where $r_{i,j}$ and $r_{j,i}$ describe the bi-directional connection between fragment feature f_i and f_j. Then, we gain the final relation matrix via transposing the original relation matrix R followed by adding it to itself, which can be written as: $\Re = R^T + R$. We feed the final relationship matrix $\Re$ into a 1×1 convolutional layer followed by BN and Softmax activation function to a relation representation †. It is obvious that every column of R is a relation representation of the fragment feature f_i. Finally, we will exploit the fragment features f_i and the corresponding relation feature to conduct dot-product for gaining the relevant global relation-aware feature matrix d: $d = \dagger \odot f_i$.

The above is a generalized representation of extracting the global relation-aware features of a certain modality via the global relation-aware network (GRN). For the specific modalities, the global relation-aware features learned by the GRN are T_r for the text and V_r that learned by the GRN for the image, which can be shown in Fig. 2.

3.4 Cross-Modal Co-Attention Network

Since the textual and visual modality have been associated in high-level semantic space, thus they can distinguish the vital feature representations separately via lining up with each other. In this part, in order to fuse the visual and textual information of given posts effectively, we use the cross-modal co-attention networks (as shown in Fig. 2) to synchronously model both the inter-modality and intra-modality relevance with their global relations for obtaining high-quality cross-modal representations of posts.

Specifically, we utilize a novel dual-stream transformer unit to handle the visual and textual information at the same time. Then, modifying the key-value query-conditioned attention mechanism to expand a multi-modal co-attention unit. In each transformer layer, given the textual and visual fragment representations $T_r = \{t_{r_1}, t_{r_2}, \ldots, t_{r_n}\}$ and $V_r = \{v_{r_1}, v_{r_2}, \ldots, v_{r_k}\}$, the module computes three matrices in each stream (i.e., Q, K, and V) relevant to queries, keys, and values on the basis of the standard transformer unit. The queries of every modality are fed into the multi-head attention module of other modalities. Then, the attention block generates visual-guided text attention in the textual stream and the text-guided visual attention in the visual stream. The rest of the multi-modal co-attention module block proceeds as the standard transformer block, including residual connection with layer normalization and a position-wise feedforward network. Then, we can obtain the image-guided textual semantic features $H_t = \{h_{t_1}, \ldots, h_{t_n}\}$ and the text-guided visual semantic features $H_v = \{h_{v_1}, \ldots, h_{v_n}\}$. Subsequently, we feed H_t into Text-CNN for the words fragment relation features in text, which can be written as:

$$h_{to} = Text - CNN(H_t) \tag{5}$$

And then, passing H_v into an average pooling layer for image region relation features in image, which can be denoted as:

$$h_{vo} = \frac{1}{k}\sum_{i=1}^{k} h_{v_i} \tag{6}$$

We get the cross-modal feature representation of the post via sum operation among h_{to} and h_{vo}, written as follows:

$$c = \lambda h_{to} + (1 - \lambda) h_{vo} \tag{7}$$

where $\lambda \in \{0, 1\}$ denotes the hyperparameter, balancing the ratio of visual and textual info among the cross-modal representations.

3.5 Fake News Detector

The inputs of fake news detector are cross-modal features $C = \{c_{s_1}, c_{s_2}, \ldots, c_{s_m}\}$, and its purpose is to classify the post as real or fake. It employs a fully-connected layer with a corresponding activation function to predict whether the post is fake or not, which may be described as:

$$\hat{p}_i = \sigma(W c_{s_i} + b) \tag{8}$$

where $\sigma(.)$ represents softmax function, and $\hat{p}_i$ is the classifying probability that indicates post i real or fake, and c_{s_i} is the cross-modal feature of post i. Then, employing y_i to denote the ground-truth label of post i. After that, using the cross-entropy algorithm as the objective function to compute the whole

loss of the whole model, which may be formalized as:

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} -[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \tag{9}$$

where m denotes the number of posts. Then, we seek an optimal parameter θ^* via minimizing the objective classification loss to end-to-end optimize the whole model, written as follows:

$$\theta^* = \arg\min_{\theta} L(\theta) \tag{10}$$

4 Datasets

4.1 Datasets

Our proposed approach CRAN will be compared with advanced baselines among two real-world datasets: PHEME [13] and WEIBO [6]. The source of PHEME dataset is five breaking news, and every news on this dataset includes a group of posts. WEIBO dataset is derived from Xin Hua News Agency[1] and Weibo[2]. Each post of its containing post id, text and image. Overall, every dataset has a sizable amount of images and texts along with corresponding label info. And the following Tab. 1 describes statistics of the two datasets.

Table 1: The statistics of two public real-world datasets

News	PHEME	WEIBO
# of real news	3880	4779
# of fake news	1972	4748
# of images	2672	38853

4.2 Baselines

In order to validate the performance of CRAN, we tend to compare our CRAN with two different types of state-of-the-art algorithms: single-modal and multi-modal methods.

1) SVM-TS [1]: It adopts an SVM linear classifier on the basis of heuristic rules to identify fake news.
2) CNN [2]: It employs the convolutional neural networks along with fixed-length windows on the target posts to learn the feature representations.
3) GRU [4]: It utilizes a multi-layer GRU unit to predict whether the post is fake or not via taking the text of posts as variable-length time series.
4) TextGCN [10]: It views the whole corpus as a heterogeneous graph and then passes it into the GCN module to capture the text representations.
5) att-RNN [5]: It generates embedding vectors of text along with its corresponding social context through LSTM unit. Then, it combines the collaborative representations with image representations via neural attention. Note that we remove the part dealing with the social context information to get a fair comparison.

[1] http://www.xinhuanet.com/
[2] https://weibo.com/

6) EANN [6]: It identifies the fake news via learning the event-invariant representations of every post, which uses an adversarial network to remove event-specific parts of post feature representations on the basis of the concatenation of learned image and text feature representations.
7) MVAE [7]: It designs a variational autoencoder including encoder and decoder of every modality of posts to learn the cross-modal features among image and text, and then uses a SoftMax classifier to predict.
8) SAFE [11]: It employs a neural network module to gain the latent feature representations of the posts containing text and image. Then it uses the relationship (similarity) among different modalities merged with the textual and visual representations to generate the final presentations of posts for debunking fake news.

Besides, several variants have been designed to prove the effectiveness of each component of our model (CRAN) as well. Then, introducing the details of its variants in the following part, which describes the analysis of CRAN components.

4.3 Parameter Setting

We employ Accuracy as the evaluation metric of the performance of our model. Since our task on the given datasets (PHEME and WEIBO) is a binary classification along with imbalanced data, we extra increase F1 score, Recall and Precision as the supplementary evaluation metrics of distinct classes to improve its reliability for identifying fake news.

On the experimental datasets, regardless of PHEME or WEIBO, according to 7:1:2, its data has been split into three shares: training set, validation set and test set.

In the Image and Text Feature Extractor module, for the image branch, the size of region features is $4 \times 4 \times 2048$. Besides, for the text branch, the BERT unit composes of 12 heads, 12 attention layers and 768 hidden units for every token.

In the Cross-modal Co-attention Network module, for the visual fragment feature representations, in order to reduce the dimension from 2048 to 768 to fit the task, we employ a 2D-convolutional layer. And the core Transformer unit uses 4 attention heads. In the Text-CNN unit, 256 filters are used for each filter size.

The Adam optimizer [15] is adopted for the whole model. For each dataset we adopted, regardless of WEIBO or PHEME, the learning rate is 0.0005 during the training phase for 100 epochs. And for all experiments, the batch size is 100.

4.4 Results

The detailed experimental results among all baselines compared with our CRAN on PHEME and WEIBO datasets in Tab. 2, from which the following findings can be obtained:

1) Among PHEME and WEIBO datasets, it is obvious that deep learning methods have superior performance to traditional machine learning methods. And we can find that SVM-TS performs worst, which implies that the hand-crafted features cannot identify fake news sufficiently. On the contrary, deep learning approaches, such as CNN, GRU, and TextGCN, perform better when compared with SVM-TS. Also, CNN has inferior performance than most existing baseline models. Based on the above results, we infer that CNN may not learn long-distance semantics relations among words may cause its poor performance. In addition, TextGCN has performed better over the given datasets, suggesting that the performance of the model can be improved via adopting graph convolutional networks.

2) Comparing single-modal and multi-modal methods on experimental datasets, it is clear that most multi-modal models bring about better accuracy than single-modal models, which indicates that visual information can supply some supplementary information, which is beneficial to detecting fake news. For instance, att-RNN performs better comparatively, implying that the performance of the model can be improved via introducing an attention mechanism. Besides, SAFE performs also relatively better, showing that integrating similarity representations among fragments of different modalities is effective. In all multi-modal models, MVAE has the best performance, indicating that adding self-supervised loss can enhance the generalization ability of model. Rather, EANN performs worst relatively, showing that in most cases eliminating event-specific representations has depraved the discriminative capability of features of posts.
3) The CRAN we proposed performs advantageously over the state-of-the-art baseline methods among all experimental datasets consistently, which demonstrates that CRAN can capture more precise and comprehensive relational cross-modal representations by jointly modeling the inter-modality and intra-modality relations with the global relational semantics in each modality in a unified framework, which is effective for identifying fake news.

Table 2: The results of baselines compared with CRAN on PHEME and WEIBO datasets

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1-score	Precision	Recall	F1-score
WEIBO	SVT-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	CRAN	**0.883**	0.901	0.862	0.881	0.867	0.905	0.886
PHEME	SVT-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	att-RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	CRAN	**0.885**	0.841	0.815	0.828	0.906	0.920	0.913

4.5 Ablation Analysis

Since the proposed CRAN model includes multiple key components, in this part, the variants of CRAN will make comparison with the below aspects to explain the effectiveness of CRAN:

1) **CRAN¬r**: A variant of CRAN with the global relation-aware network module being eliminated.
2) **CRAN¬c**: A variant of CRAN with the cross-modal cc-attention network module being eliminated.
3) **CRAN¬v**: A variant of CRAN with the visual information being eliminated.

We conduct the ablation analysis between our model and its variants, and the ablation results are shown in Tab. 3, from which we can obtain the following findings:

1) **Effects of global relation-aware network**: We contrast the performance of CRAN and CRAN¬r among the experimental datasets. It is obvious that CRAN has performed better than CRAN¬r, which confirms the superiority of introducing the global relation-aware network to our model, which explores the global relation semantic information of the target modality of posts.
2) **Effects of cross-modal co-attention network**: We contrast the performance of CRAN with CRAN¬c on PHEME and WEIBO datasets. It is clear that CRAN has performed more superior to CRAN¬c, which confirms the superiority of introducing the multi-modal co-attention network to our model.
3) **Effects of the visual information**: We contrast the performance of CRAN and CRAN¬v over the experimental datasets. We can find that CRAN has performed more excellently than CRAN¬v, which indicates that visual information can supply supplementary semantic info to enhance the performance of our model.

Table 3: The results of CRAN compared with its variants on PHEME and WEIBO datasets

Datasets	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1-score	Precision	Recall	F1-score
WEIBO	CRAN¬r	0.866	0.881	0.847	0.863	0.852	0.885	0.868
	CRAN¬c	0.829	0.779	0.921	0.844	0.903	0.737	0.812
	CRAN¬v	0.794	0.844	0.721	0.778	0.756	0.867	0.807
	CRAN	**0.883**	0.901	0.862	0.881	0.867	0.905	0.886
PHEME	CRAN¬r	0.878	0.832	0.803	0.817	0.900	0.916	0.908
	CRAN¬c	0.826	0.780	0.681	0.727	0.846	0.901	0.872
	CRAN¬v	0.855	0.811	0.749	0.779	0.876	0.910	0.892
	CRAN	**0.885**	0.841	0.815	0.828	0.906	0.920	0.913

4.6 Parameter λ Analysis

In this part, we will discuss the effectiveness of the hyper-parameter λ, whose function is to balance the proportion of visual information and textual information in the posts. We will experiment among the two real-world public datasets (PHEME and WEIBO), respectively, whose results will be described in Fig. 3. During the experiment stage, parameter λ can be viewed as a variable, where $\lambda \in \{0, 1\}$. In order to seek suitable parameter λ, we set $\lambda = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and use the Accuracy and F1 Score as evaluation metrics to weigh the capability of the model under the impacts of this parameter λ, where the value of Accuracy ranges 0% to 100%, and the value of F1 Score is set to range from 0.0 to 1.0.

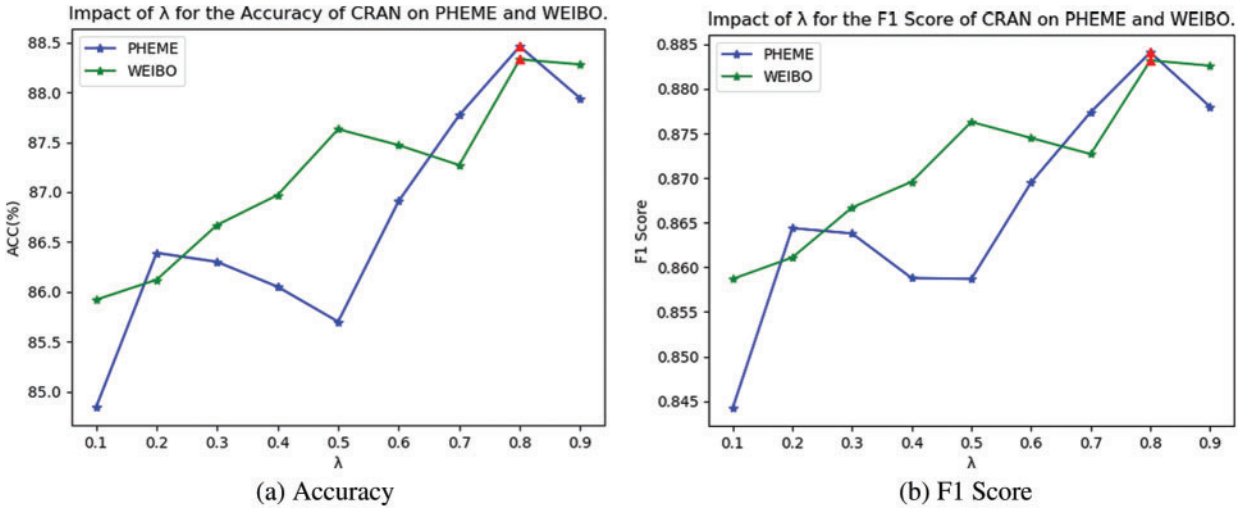

Figure 3: Impacts of λ for the accuracy and F1 score of CRAN on PHEME and WEIBO datasets

From Fig. 3a, when λ gets the value from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, we can see that the results are like for the Accuracy metrics of the performance of CRAN on the two datasets. That is to say, when $\lambda = 0.8$, the Accuracy can get the maximum regardless of PHEME or WEIBO datasets. Simultaneously, from Fig. 3b, we can see that with the change of the value of λ, where $\lambda = 0.8$, whether it is on PHEME or WEIBO datasets, F1 Score gets the highest.

To sum up, we set λ as 0.8, and based on this setting, our CRAN model can achieve the optimal performance among the two real-world public datasets (PHEME and WEIBO).

5 Conclusion

In this paper, we propose an innovative Cross-modal Relation-aware Networks (CRAN) for performing fake news detection task via jointly modeling the textual and visual information with their global relations in a unified end-to-end model. We discuss that most existing approaches cannot explicitly model the global binary semantic relations among image regions or text tokens to learn discriminative representations in each modality regardless of image or text. Besides, most existing methods cannot exploit the supplementary multi-modal information including image and text relations to supplement and enrich each modality effectively. To handle the above limitations, CRAN has been proposed to introduce two innovations: 1) design a global relation-aware network for each modality of the social multimedia posts to mine the deep semantic representations among fragments containing the global scope relation information of the target modality such as text and image. 2)

propose a novel cross-modal co-attention network module for cross-modal information fusion, which is capable of utilizing the intra-modality and inter-modality relationships jointly along with their inner global relation information among image and textual fragments to supplement and heighten each other for high-quality cross-modal representation. Since methods based on background knowledge of the content of posts cannot be fully researched, in future work, our aim is to explore an effective means to employ background knowledge by exploiting deep neural networks, which supplies more beneficial supplementary information to detect fake news in real scenes.

Funding Statement: This paper is partially funded by the National Natural Science Foundation of China (Grant No. 61902193); and in part by the PAPD fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Ma, W. Gao, Z. Wei, Y. Lu and K. -F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. of the 24th ACM International on Conference on Information and Knowledge Management*, New York, pp. 1751–1754, 2015.

[2] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan *et al.,* "A convolutional approach for misinformation identification." in *IJCAI*, Melbourne, pp. 3901–3907, 2017.

[3] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[4] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. -F. Wong and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," 2016.

[5] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, no. Nov, pp. 45–66, 2001.

[6] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. of the 25th ACM International Conf. on Multimedia*, Mountain View, CA USA, pp. 795–816, 2017.

[7] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun *et al.,* "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proc. of the 24th acm Sigkdd International Conf. on Knowledge Discovery & Data Mining*, London, pp. 849–857, 2018.

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

[9] X. Zhou, J. Wu and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, San Diego, Springer, pp. 354–367, 2020.

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

[11] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, USA, pp. 770–778, 2016.

[12] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2–7, vol. 1 (Long and Short Papers), In: J. Burstein, C. Doran and T. Solorio (Eds.). Association for Computational Linguistics, 2019, pp. 4171–4186, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1423.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108.

[14] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," in *Chinese Computational Linguistics-18th China National Conf., CCL 2019*, Kunming, China, October pp. 18–20, 2019, Proceedings, ser. Lecture Notes in Computer Science, In: M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu (Eds.), vol. 11856, Springer, pp. 194–206, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-32381-3_16.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.