Cross-modal Attention Network with Orthogonal Latent Memory for Rumor Detection

Zekai Wu[1], Jiaxin Chen[1], Zhenguo Yang[1(✉)], Haoran Xie[2], Fu Lee Wang[3], and Wenyin Liu[1,4(✉)]

[1] Guangdong University of Technology, Guangzhou, China
{yzg,liuwy}@gdut.edu.cn
[2] Lingnan University, Hong Kong, China
hrxie@ln.edu.hk
[3] Hong Kong Metropolitan University, Hong Kong, China
pwang@hkmu.edu.hk
[4] Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen, China

Abstract. In this paper, we design a cross-modal attention fusion network with orthogonal latent memory (CALM) to fuse multi-modal social media data for rumor detection. Given multimodal content features extracted from text and images, we devise a cross-modal attention fusion (CAF) mechanism to extract critical information underlying the modalities by intra-modality attention, and model the underlying relations among the modalities by inter-modality attention. In terms of the text, the natural sequential characteristics are critical to semantic understanding, while existing sequence models suffer from losing the information conveyed by the former words. To this end, we propose a Bi-GRU with orthogonal latent memory to extract the sequential features from the text, where the memory captures independent patterns. The fused content features and the sequential features can be used for rumor detection seamlessly. Extensive experiments conducted on two real-world datasets show the outperformance of the proposed CALM. (e.g., F_1-score is improved from 0.823 to 0.846 on Weibo dataset).

Keywords: Rumor detection · Multi-modal · Social media

1 Introduction

Social media has revolutionized the way for people to acquire information, while it may foster the propagation of fake news like rumors in turn. Some offenders even use rumors to guide public opinion, damage the credibility of the government and even interfere with the general election [1]. Rumor detection aims to identify the rumors distributed on social media like platforms, where the data usually are in multiple modalities such as text, image, and videos, etc., being verisimilitude to the interest of most people.

W. Zhang et al. (Eds.): WISE 2021, LNCS 13080, pp. 527–541, 2021.
https://doi.org/10.1007/978-3-030-90888-1_40

In terms of the methodologies, the early works usually focus on textual news. For instance, Castillo et al. [4] extract message-based and topic-based features from the textual content and exploit a decision tree method to classify posts. Yu et al. [16] use a convolutional approach to extract key features and shape high-level interactions from textual content of the relevant posts. Recent studies have shown that detecting rumors in a multi-modal manner can achieve better performance, especially with deep learning methods. For instance, Khattar et al. [7] propose a novel VAE model to learn a shared representation of the modalities for detecting rumors. Yang et al. [15] apply a Ti-CNN method to detect rumors by extracting both explicit and latent multi-modal features within news content. In terms of fusing the heterogeneous modalities in the context of rumor detection, quite a few fusion strategies show impressive performance. For instance, Jin et al. [6] propose an attention mechanism to fuse visual, textual and social context features. Chen et al. [5] propose a self-attentive fusion mechanism to integrate the textual features with visual features. The aforementioned approaches suffer from a few deficiencies. Firstly, these methods either pay more attention to the semantic information or sequential information in social media textual data merely. Secondly, the existing approaches usually concatenate the multimodal features or introduce attention mechanism to weight the importance of modalities, neglecting correlations and interactions underlying the modalities.

In this paper, we design a cross-modal attention fusion network with orthogonal latent memory (denoted as CALM) to detect rumors from multimodal social media data. On one hand, we propose a cross-modal attention fusion mechanism with intra-modality and inter-modality attentions, where intra-modality attention extracts critical information underlying the single modalities, and inter-modality attention establishes the relations among multiple modalities. On the other hand, we extend Bi-GRU with orthogonal latent memory to capture long-distance temporal dependencies in the sequential models, avoiding gradient vanishing and exploding. In particular, orthogonal constraint on the latent memory ensures the diversity of the underlying patterns from global viewpoint.

The main contributions are summarized as follows.

- We propose a cross-modal attention fusion framework with intra-modality and inter-modality attentions to capture the modality-specific information and model the underlying relations among the multiple modalities.
- We devise an orthogonal latent memory to keep diverse latent patterns from the global viewpoint, which can be plugged in GRU-like sequential models to capture the long-distance temporal dependencies.
- We conduct extensive experiments on two real-word datasets, which show the outperformance of the proposed approach compared with the state-of-the-art baselines.

The rest of this paper is organized as follows. Section 2 summarizes the related works. Section 3 presents the proposed CALM. Section 4 shows the experiments and analyzes the experimental results. Section 5 concludes the work.

2 Related Work

In this section, we briefly review the works on multi-modal rumor detection and multi-modal data fusion.

2.1 Multi-modal Rumor Detection

Social media has become the main platform for people to obtain and share information, which may lead to the spread of rumors extremely fast in turn. The research attention on rumor detection has shifted from text-based approaches to multi-modal ones recently. For instance, Zhang et al. [18] employed a pre-trained BERT model to identify rumors and used a domain classifier to remove event-specific dependency. Zhang et al. [17] designed a knowledge-aware network and an event memory network for social media rumors. Zhou et al. [19] exploited multi-modal and relational information to learn the representation of articles and predict rumors. However, the textual extractor employed by prior studies either mainly focused on the semantic information or sequential information.

2.2 Multi-modal Data Fusion

Multi-modal data fusion aims to combine multi-aspect information from multiple data modalities, which are critical for various machine learning tasks [8,10]. In the context of rumor detection, quite a few multi-modal data fusion approaches have been devised to deal with the multimodal data. For instance, Wang et al. [14] concatenated the visual features and textual features of social media data to get a multi-modal feature. Jin et al. [6] proposed a recurrent neural network with an attention mechanism to fuse image and text features. Chen et al. [5] proposed a self-attentive fusion mechanism to integrate the textual features with visual features for detecting rumors. The aforementioned methods can hardly discover latent correlations among the multiple modalities as the complementarity among multimodal features has not been fully explored.

3 Methodology

3.1 Overview of the Framework

The overall framework of the proposed CALM is shown in Fig. 1, which consists of four components, i.e., the visual extractor, the textual extractor, the cross-modal attention fusion (CAF) network and the rumor detector. The visual extractor and textual extractor extract visual and textual features from social media data. Specifically, the textual extractor can extract both semantic features and sequential features. Furthermore, the CAF component fuses multimodal content features extracted from text and images by inter-modality attention and intra-modality attention. Finally, the rumor detector concatenates the learned features as input to predict whether the social media data is rumor or non-rumor.

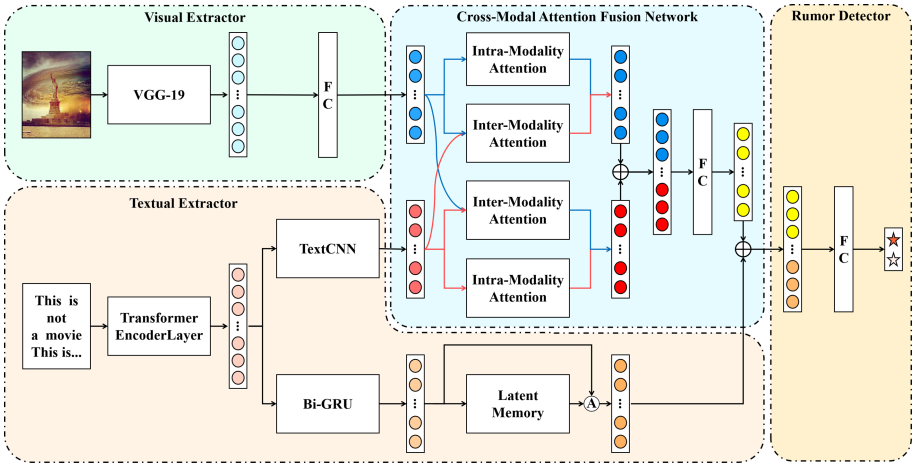


Fig. 1. Overview of CALM.

3.2 Visual Extractor

The attached image v of the social media data is fed into the visual extractor. We employ the VGG-19 network [11] to extract visual features, which has achieved impressive performance on multiple computer visual tasks. We extract the CNN feature for image from the fc-7 layer in VGG-19 and feed it into a fully connected layer to reduce the dimension down to d_m. The visual features $V \in R^{d_m}$ can be obtained as follows:

$$V = \sigma\left(W_v \cdot VGG\left(v\right)\right) \tag{1}$$

where $VGG\left(\cdot\right)$ is the pre-trained VGG-19 model, W_v is the weight matrix of the fully connected layer and $\sigma\left(\cdot\right)$ is the activation function used.

3.3 Textual Extractor

We divided textual extractor into two sub-modules, content feature extraction and sequential feature extraction.

1) Content Feature Extraction. The textual input t to the textual extractor is the sequential list of words in the posts, $t = [t_1 t_2 \cdots t_n]$, where n is the number of words in the text. Each word $t_i \in t$ is represented as a word embedding vector, which is extracted with a pre-trained word2vec model. In order to obtain better understanding of the language structure, we employ the Transformer Encoder [12] to calculate and assign weights for different words in t. With E denotes as the encoder output, the operation can be obtained as follows:

$$E = TransformerEncoder\left(t\right) \tag{2}$$

noted that $E = [E_1 E_2 \cdots E_n]$, where E_i is the encoder result of t_i.

More specifically, to capture the semantic features from text of the social media data, the content feature extraction exploits Text-CNN [9] to automatically capture semantic features in different granularities. Furthermore, the feature map produced by Text-CNN is fed into a fully connected layer to ensure the semantic features have the same dimension as the visual features V. Given the encoder output E, the semantic features $T \in R^{d_m}$ can be calculated as follows:

$$T_t = TextCNN\left(E\right) \tag{3}$$

$$T = \sigma\left(W_t \cdot T_t\right) \tag{4}$$

where $TextCNN\left(\cdot\right)$ is the Text-CNN model and W_t is the weight matrix in the fully connected layer.

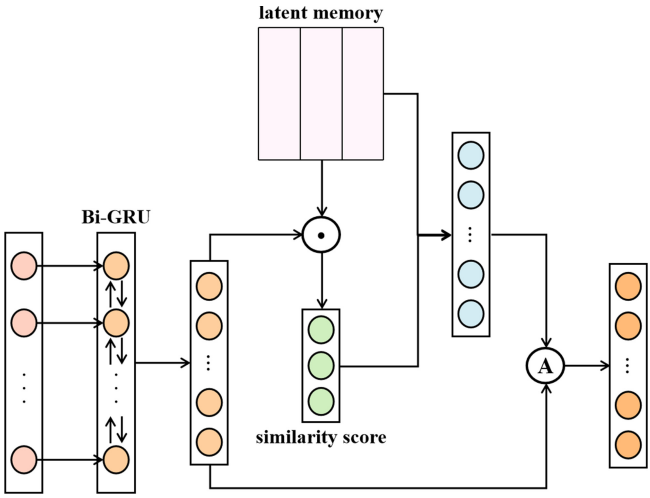


Fig. 2. The details of sequential feature extraction.

2) Sequential Feature Extraction. Existing sequence models suffer from a problem of vanishing and exploding gradients. This leads to the model learning inefficient dependencies between words that are a few steps apart. To overcome this problem, a latent memory network is introduced to improve Bi-GRU, which can not only make up for the defects of the sequence models, but also output the extra global latent patterns information shared by rumors. The details of the sequential feature extraction are provided in Fig. 2.

More specifically, given the input E, we use a Bi-GRU to compute the hidden state for each element and concatenate the last hidden state from both directions, denoted as $R_{gru} \in R^{2*d_m}$. Subsequently, we pass the R_{gru} through a fully connected layer to calculate the preliminary sequence features F_g. The operation can be represented as follows:

$$R_{gru} = GRU_{bi}\left(E\right) \tag{5}$$

$$F_g = \sigma\left(W_g \cdot R_{gru}\right) \tag{6}$$

where $GRU_{bi}(\cdot)$ represents the Bi-GRU model and W_g is the weight matrix of the fully connected layer.

Furthermore, the patterns information in memory are chosen to strengthen the sequence features. In particular, the memory network is denoted as $M \in R^{num \times d_m}$, where num is depended on the number of latent patterns underlying the social media data. We calculate the similarity score M_{score} between the sequence features F_g and the latent patterns, which can be obtained by conducting softmax function on their dot product as follows:

$$M_{score} = softmax\left(M^T \cdot F_g\right) \tag{7}$$

Finally, we extract the closest patterns based on the similarity score and merge the resulting patterns information F_m with the sequence features F_g through conducting average operation. The final sequence features $T_g \in R^{d_m}$ can be obtained as follows:

$$F_m = (M \cdot M_{score}) \tag{8}$$

$$T_g = avg\left(F_g,\ F_m\right) \tag{9}$$

where $avg(\cdot)$ represents the average operation.

3.4 Cross-modal Attention Fusion Network (CAF)

In terms of multi-modal feature fusion, the visual features and semantic features are extracted by different methods, meaning it is not suitable to concatenate

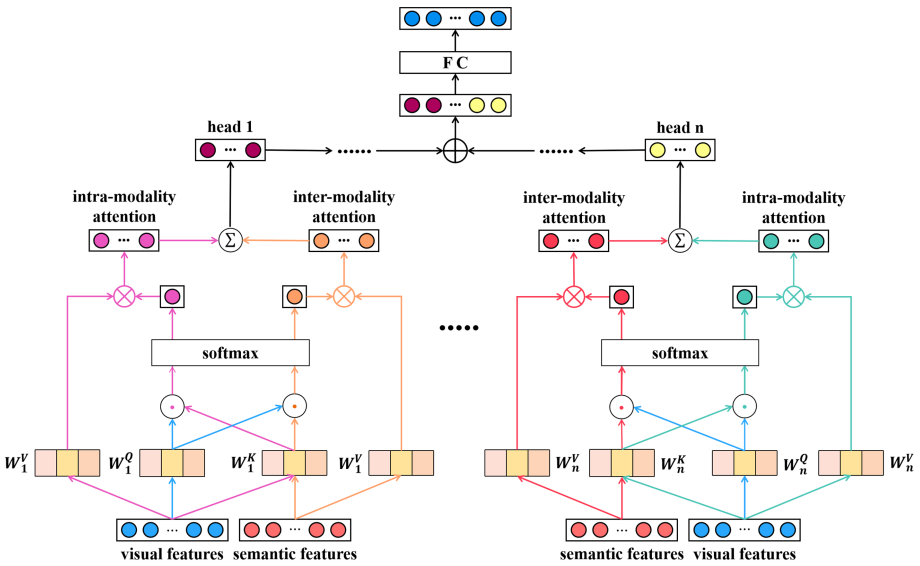


Fig. 3. The architecture of CAF.

them together directly. To this end, we devise a cross-modal attention fusion mechanism with intra-modality and inter-modality attentions to improve traditional fusion strategy. As shown in Fig. 3, each modality should not only pay attention to its own characteristics but also focus on other modal features. In particular, the multi-head mechanism allows the CAF to extract information from different feature spaces, which help the model explore different attention patterns in a variety of angles.

1) Intra-Modality Attention. Given the multimodal content features, we first produce a set of query, key and value pair by linear transformations for single modality. Taking the visual features V as an example, the operation can be obtained as follows:

$$V_Q = Linear(V,\ W^Q) \tag{10}$$

$$V_K = Linear(V,\ W^K) \tag{11}$$

$$V_V = Linear(V,\ W^V) \tag{12}$$

where 'Linear" denotes a fully connected layer, $W^Q, W^K, W^V \in R^{d_m \times d_h}$ are the weight matrices and d_h represents the common dimension of the transformed features obtained from multiple modalities. Similarly, the corresponding linear transformations for the semantic features T can be represented as T_Q, T_K and T_V.

More specifically, we calculate the scaled dot product as the intra-modality attention weight. Given the V_Q and V_K, the operation can be obtained as follows:

$$V_{intra} = \frac{\left(V_Q \cdot {V_K}^T\right)}{\sqrt{d_h}} \tag{13}$$

where V_{intra} represents the intra-modality attention weight for V. Correspondingly, the semantic intra-modality attention weight T_{intra} can be calculated as follows:

$$T_{intra} = \frac{\left(T_Q \cdot {T_K}^T\right)}{\sqrt{d_h}} \tag{14}$$

2) Inter-modality Attention. As for inter-modality attention, to model the underlying relations among multiple modalities, we learn the inter-modality attention weight in a similar way.

$$V_{inter} = \frac{\left(V_Q \cdot {T_K}^T\right)}{\sqrt{d_h}} \tag{15}$$

where V_{inter} represents the inter-modality attention weight for V.

Furthermore, the softmax function is used to normalize the intra-modality and inter-modality attention weights. Then the visual resulting features V_C can be obtained by weighted summation over the different modalities.

$$V_C = softmax\left([V_{intra}, V_{inter}]\right) \begin{bmatrix} V_V \\ T_V \end{bmatrix} \tag{16}$$

In particular, the CAF calculates the intra-modality and the inter-modality attentions h times respectively and concatenates the multi-head features together. For clarity, we define that V_{C_i} is the attention outcome in the i^{th} head and W_i^Q, W_i^K, W_i^V are the weight matrices used in the corresponding linear transformations. In addition, we exploit a weight matrix to reduce the dimension for each modality. The operation can be obtained as follows:

$$F_v = W_o \cdot [V_{C_1} \oplus V_{C_2} \oplus \cdots \oplus V_{C_h}] \tag{17}$$

where $\oplus$ denotes the concatenation operation, $W_o \in R^{h*d_h \times d_m}$ is the weight matrix and F_v represents the visual resulting features obtained from the cross-modal attention.

Relatively, the cross-modal attention outcome for the semantic features T can be achieved in a similar way, which is denoted as F_t.

$$F_t = W_o \cdot [T_{C_1} \oplus T_{C_2} \oplus \cdots \oplus T_{C_h}] \tag{18}$$

Finally, we concatenate multimodal resulting features together and exploit a fully connected layer to calculate the final fused content features $T_f \in R^{d_m}$ as follows:

$$T_f = \sigma\left(W_f \cdot (F_v \oplus F_t)\right) \tag{19}$$

where W_f is the weight matrix of the fully connected layer.

3.5 Rumor Detector

The goal of the rumor detector is to identify whether a social media data is a rumor or non-rumor. Given the fused content features T_f and the sequence features T_g, the rumor detector concatenates above features seamlessly and feeds the features into two fully connected layers to output the predicted result $\tilde{y}$. The operation of the detector can be represented as follows:

$$\tilde{y} = softmax\left(W_{r2} \cdot \sigma\left(W_{r1} \cdot (T_f \oplus T_g)\right)\right) \tag{20}$$

where W_{r1}, W_{r2} are the weight matrices of the fully connected layers.

3.6 Loss Function

In terms of loss function used, we design an orthogonal constraint to make the latent memory keep its orthogonality and exploit a rumor detection loss function to identify rumors.

1) Orthogonal Constraint. The orthogonal constraint aims to minimize the pairwise cosine similarity between the patterns in the latent memory, which ensures the variety of the patterns to improve the discriminative power of the

memory. More specifically, given the latent memory network M, the proposed constraint can be represented below:

$$C_\beta(M) = \beta \left\| M^T M \odot (1 - I) \right\|_F^2 \tag{21}$$

where 1 denotes a matrix with all elements set to 1, $\odot$ represents the element-wise product, I is the identity matrix and β is a hyperparameter.

2) Rumor Detection Loss. To identify rumors, we define a loss term $\mathcal{L}$ by using cross entropy as follows:

$$\mathcal{L} = \sum_{i}^{N} -[y_i \times \log(\tilde{y_i}) + (1 - y_i) \times \log(1 - \tilde{y_i})] \tag{22}$$

where $\tilde{y_i}$ is the predicted result obtained from rumor detector for the i^{th} sample, and y_i is the corresponding ground-truth. N is the total number of social media samples.

Finally, the loss function of CALM can be written as follows:

$$\mathcal{L}_{CALM}(\theta, M) = \mathcal{L} + C_\beta(M) \tag{23}$$

where θ is denoted as the parameter set of the proposed CALM.

The detailed steps of the proposed model CALM are summarized in Algorithm 1.

Algorithm 1. The **CALM** algorithm

Input: label $y = \{y_i\}_{i=1}^N$, textual input $t = \{t_i\}_{i=1}^N$, visual input $v = \{v_i\}_{i=1}^N$, β for orthogonal constraint, the latent memory M, learning rate lr.

Initialize the model parameters
Set the status of the model for training
for number of training iterations **do**
  $T, T_g = \text{TextualExtractor}(t, M)$
  $V = \text{VisualExtractor}(v)$
  $T_f = \text{CAF}(T, V)$
  $\tilde{y} = \text{RumorDetector}(T_f \oplus T_g)$
  Compute the loss using the loss function $\mathcal{L}_{CALM}(\theta, M)$ with $\tilde{y}$ and y
  Decay learning rate lr according to the number of the training iterations
end for

4 Experiments

4.1 Datasets

1) Twitter Dataset [3], is comprised of 514 images and 18,264 Tweets. We filter out the Tweets with noise and unclear labels, resulting in 379 images and 15,629 Tweets being related 9,405 rumors and 6,224 non-rumors.

2) Weibo Dataset [6], consists of 9,527 posts being related 4,748 rumors and 4,779 non-rumors. We split the dataset into training, validation, and testing sets in a ratio of 7:1:2.

4.2 Baselines

We compare the proposed methods with the following baselines:

1) VQA [2], aims to answer the questions about the given images. We improve the original VQA model to adapt to the rumor detection.
2) NeuralTalk [13], averages the outputs of RNN at each time step to obtain the latent representations and generates corresponding description for the given images.
3) att-RNN [6], uses the attention mechanism to fuse the visual, textual and social context features for rumor detection.
4) EANN [14], designs three components for multimodal rumor detection, including multimodal feature extractor, fake news detector and event discriminator.
5) MVAE [7], devises a multi-modal VAE structure to obtain shared representation and employs a binary classifier to detect rumors.
6) MFN [5], exploits a self-attentive mechanism to integrate multi-modal information and introduces a latent topic network to detect upcoming rumors.
7) BDANN [18], employs a BERT-based approach to extract multi-modal features and proposes a domain classifier to remove the event-specific dependency. As the domain classifier requires event labels, for a fair comparison, we remove the domain classifier in BDANN.

In terms of evaluations, accuracy, precision, recall, and F_1-score are adopted.

4.3 Performance of the Approaches

Table 1 summarizes the performance of the approaches on two datasets, from which we have some observations. 1) The multi-modal rumor detection models, e.g., att-RNN, EANN and CALM, outperform the multimodal fusion methods for rumor detection, such as VQA and NeuralTalk. The reason may be that the rumor detection models make full use of information about rumor and non-rumor events, e.g., global latent rumor patterns, event information, etc. 2) In terms of the rumor detection approaches, CALM significantly outperforms the baselines, benefiting from the cross-modal attention fusion mechanism to integrate multi-modal information and the orthogonal latent memory to capture robust representations.

Table 1. Performance of the approaches on two datasets

Dataset	Method	Accuracy	Rumors			Non-Rumors		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	VQA	0.753	0.719	0.601	0.655	0.769	0.849	0.807
	NeuralTalk	0.667	0.570	0.593	0.582	0.733	0.714	0.723
	att-RNN	0.756	0.724	0.604	0.658	0.771	0.853	0.810
	EANN	0.757	0.728	0.601	0.658	0.770	0.856	0.811
	MVAE	0.805	0.869	0.588	0.702	0.782	**0.943**	0.855
	MFN	0.808	0.850	0.616	0.715	0.791	0.931	0.855
	BDANN	0.827	**0.872**	0.652	0.746	0.808	0.939	0.869
	CALM	**0.845**	0.785	**0.831**	**0.807**	**0.888**	0.855	**0.871**
Weibo	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	NeuralTalk	0.726	0.794	0.613	0.692	0.684	0.840	0.754
	att-RNN	0.788	**0.862**	0.686	0.764	0.738	**0.890**	0.807
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MFN	0.803	0.811	0.806	0.808	0.794	0.800	0.797
	BDANN	0.814	0.800	0.860	0.830	0.840	0.760	0.800
	CALM	**0.846**	0.843	**0.864**	**0.853**	**0.851**	0.828	**0.839**

4.4 Ablation Study

Table 2. Performance of the variations of CALM

Dataset	Method	Accuracy	Precision	Recall	F_1
Twitter	CALM_CA	0.812	0.807	0.793	0.798
	CALM_LM	0.824	0.815	0.813	0.814
	CALM_OC	0.827	0.824	0.840	0.824
	CALM	**0.845**	**0.836**	**0.843**	**0.839**
Weibo	CALM_CA	0.821	0.821	0.822	0.821
	CALM_LM	0.824	0.832	0.826	0.823
	CALM_OC	0.831	0.832	0.832	0.831
	CALM	**0.846**	**0.847**	**0.846**	**0.846**

CALM consists of a cross-modal attention fusion (CAF) mechanism to combine multimodal content features and an orthogonal latent memory network to keep diverse latent patterns. For clarity, let CALM_CA denote CALM without CAF module and CALM_LM denote CALM without latent memory network. Furthermore, we remove orthogonal constraint to evaluate the effectiveness of preserving orthogonality among latent patterns, which is denoted as CALM_OC. The performance of the variations of CALM are summarized in Table 2, from which we have the following observations. 1) CALM with all the components achieves the best performance on both datasets, demonstrating the significance of each

module. 2) We can observe that the performance of CALM drops dramatically without the CAF module. The reason is that the CAF module extracts critical information underlying the single modalities by intra-modality attention, and models the underlying relations among the modalities by inter-modality attention.

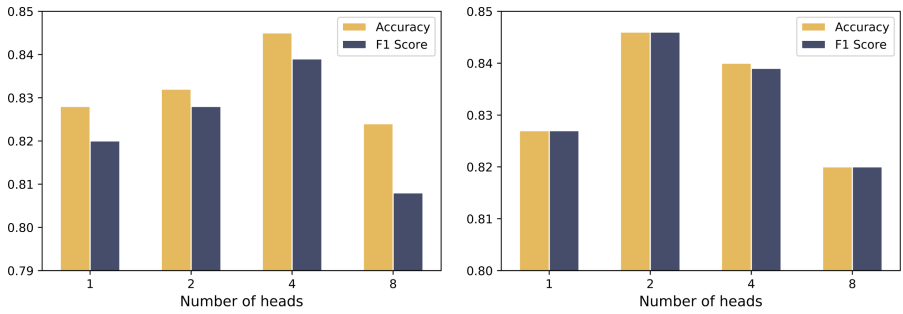
4.5 Effectiveness of CALM on Multimodal Fusion

Table 3. Performance of CALM using single or multiple modalities

Dataset	Method	Accuracy	Precision	Recall	F_1
Twitter	CALM_T	0.786	0.804	0.815	0.786
	CALM_V	0.775	0.781	0.725	0.737
	CALM	**0.845**	**0.836**	**0.843**	**0.839**
Weibo	CALM_T	0.816	0.819	0.818	0.816
	CALM_V	0.580	0.586	0.573	0.560
	CALM	**0.846**	**0.847**	**0.846**	**0.846**

Table 3 summarizes the performance of CALM using single or multiple modalities on both datasets, from which we have two-fold observations. 1) CALM uses text and image jointly outperforms it uses either text (CALM_T) or image (CALM_V) merely, indicating the necessity of multimodal fusion. 2) In terms of the single data modality, text is more effective than images as text conveys certain semantic information that is easy to understand by humans or machines.

4.6 Impact of the Number of Heads in CAF

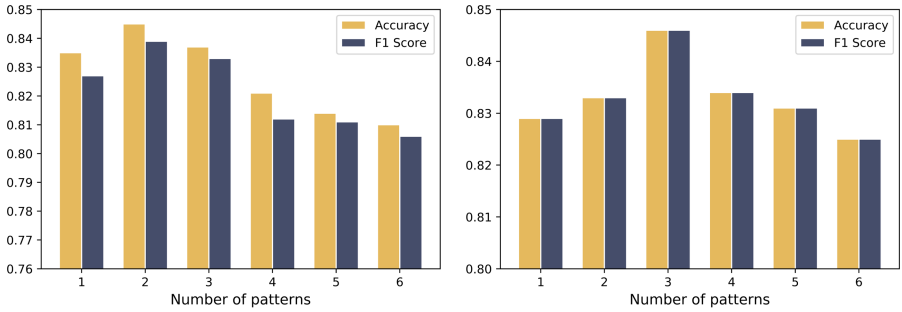


(a) Twitter (b) Weibo

Fig. 4. Impact of the number of heads in CAF.

Figure 4 summarizes the performance of CALM with different number of heads in the CAF module. We can find that maintaining a large number of heads may not necessarily improve the performance on both datasets. The reason could be that a large number of heads will increase the complexity of the model, while they cannot capture more attentional patterns than the certain number underlying the datasets.

4.7 Impact of the Number of Patterns in Latent Memory



(a) Twitter (b) Weibo

Fig. 5. Impact of the number of patterns in latent memory.

Figure 5 summarizes the performance of CALM with different number of patterns in the latent memory. We can observe that the performance of CALM can be improved with the increasing number of patterns at the beginning, while too many patterns will lead to poor result on both datasets. This is probably because that the number of latent patterns underlying the datasets is limited.

4.8 Failure Cases Study

Figure 6 shows some examples that are predicted falsely by CALM, from which we have some observations. 1) In terms of non-rumors predicted as rumors by CALM, we can observe that the images have not shown discriminative information and the textual descriptions seem to exaggerate the facts more or less. 2) In terms of rumors predicted as non-rumors by CALM, we can observe that the textual and visual contents are quite consistent and relevant, which may confuse the model and it is even hard for humans to make identification.

Fig. 6. Failure cases of CALM. The Tweets in orange background shows non-rumors which are predicted as rumors, and the Tweets in blue background shows rumors that are recognized as non-rumors. (Color figure online)

5 Conclusion

In this paper, we propose a cross-modal attention fusion network with orthogonal latent memory for rumor detection. Specifically, we exploit a cross-modal attention mechanism with intra-modality and inter-modality attentions to integrate the modality-critical information and fully explore potential hidden correlations among the modalities. In particular, the proposed network introduces an orthogonal latent memory to store global latent patterns information shared by the rumor events, which can improve sequential models to capture the long-distance temporal dependencies. The experiments conducted on two popular datasets show the effectiveness of the proposed CALM for rumor detection.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 62076073), the Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515010616), Science and Technology Program of Guangzhou (No. 202102020524), the Guangdong Innovative Research Team Program (No.2014Z T05G157), HKIBS Research Program Grant Application (HCRG-201-002) and the Faculty Research Grant (DB21B6) of Lingnan University, Hong Kong.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–36 (2017)

2. Antol, S., et al.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
3. Boididou, C., et al.: Verifying multimedia use at mediaeval 2015. Media Eval. **3**(3), 7 (2015)
4. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web (2011)
5. Chen, J., Wu, Z., Yang, Z., Xie, H., Wang, F.L., Liu, W.: Multimodal fusion network with latent topic memory for rumor detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)
6. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia (2017)
7. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: Mvae: Multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference (2019). https://doi.org/10.1145/3308558.3313552
8. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: IEEE International Conference on Acoustics (2013)
9. Kim, Y.: Convolutional neural networks for sentence classification (2014)
10. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
12. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
13. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
14. Wang, Y., et al.: Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 849–857. KDD '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3219819.3219903
15. Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: Ti-cnn: Convolutional neural networks for fake news detection (2018)
16. Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T.: A convolutional approach for misinformation identification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 3901–3907 (2017). https://doi.org/10.24963/ijcai.2017/545
17. Zhang, H., Fang, Q., Qian, S., Xu, C.: Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the 27th ACM International Conference on Multimedia (2019). https://doi.org/10.1145/3343031.3350850
18. Zhang, T., et al.: Bdann: bert-based domain adaptation neural network for multimodal fake news detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2020). https://doi.org/10.1109/IJCNN48605.2020.9206973
19. Zhou, X., Wu, J., Zafarani, R.: Safe: similarity-aware multi-modal fake news detection. Adv. Knowl. Discovery Data Mining **12085**, 354 (2020)