Multimodal Fusion with Co-Attention Networks for Fake News Detection

Yang Wu [1,2], Pengwei Zhan[1,2], Yunjian Zhang[1,2], Liming Wang[1]*, Zhen Xu[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{wuyang0419,zhanpengwei,zhangyunjian,wangliming,xuzhen}@iie.ac.cn

Abstract

Fake news with textual and visual contents has a better story-telling ability than text-only contents, and can be spread quickly with social media. People can be easily deceived by such fake news, and traditional expert identification is labor-intensive. Therefore, automatic detection of multimodal fake news has become a new hot-spot issue. A shortcoming of existing approaches is their inability to fuse multimodality features effectively. They simply concatenate unimodal features without considering inter-modality relations. Inspired by the way people read news with image and text, we propose a novel Multimodal Co-Attention Networks (MCAN) to better fuse textual and visual features for fake news detection. Extensive experiments conducted on two real-world datasets demonstrate that MCAN can learn inter-dependencies among multimodal features and outperforms state-of-the-art methods.

1 Introduction

The rapid growth of social media has created fertile soil for the emergence and fast spread of fake news (Zhao et al., 2015), resulting in serious consequences. For example, during the U.S. 2016 presidential election, the most popular fake news was more widely spread than the most popular authentic news on Facebook, which confused people and broke the authenticity balance of the news ecosystem (Shu et al., 2017). To mitigate the negative effects caused by fake news, it is crucial to detect fake news on social media automatically.

Tweets with images are getting popular on social media recently, which have richer information and attract more viewers than tweets with only texts (Jin et al., 2017). Fake news also makes full use of this advantage to draw and mislead readers. Figure 1 shows three examples of fake news from Twitter.

Sharks in the mall! After the hurricane sandy!

Lenticular Clouds over Mount Fuji.

Woman, 36, gives birth to 14 children from 14 different fathers.

Figure 1: Some fake news from Twitter.

In the left example, both text and image indicate it is likely to be fake. The text of the middle one provides little evidence that it is fake news, but the image is obviously forged. In the right example, the image seems normal, while the textual contents indicate that it is probably fake. A hypothesis drawn from these examples is that combining text and the attached image is more conducive to detecting fake news.

Recent works have a growing interest in using multimodal (text + image) information to detect fake news. Jin et al.(2017) utilize local attention mechanisms to fuse features of image, text, and social context. Some studies explore to learn the joint representations of text and image, based on auxiliary adversarial networks (Wang et al., 2018) and variational autoencoders (Khattar et al., 2019). Nevertheless, they are not fine-grained enough in feature extraction and feature fusion. First, some studies require labor-intensive extra information, such as social context (Jin et al., 2017) and event category (Wang et al., 2018), to help detect fake news, which increases the cost of the detection. Second, except for texts in tweets, the methods mentioned above all focus on characteristics of images at the semantic level (e.g., emotional provocations), which can be reflected in the spatial domain. However, these methods ignore the individual information of fake images at the physical level, e.g., re-compression artifacts, which is reflected in the frequency domain (Qi et al., 2019). Third, some

*Corresponding Author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2560–2569

models (Wang et al., 2018; Khattar et al., 2019) obtain fused representations by simply concatenating multi-modality features. Although leverages local attention mechanism, the attention values of att-RNN (Jin et al., 2017) are only obtained from joint textual-social representations, which cannot reflect the similarity between textual-social representations and visual representations. Intuitively, when people judge news credibility with text and image, they often observe image first and then read text (Wang et al., 2020). This process may be repeated several times. In this process, people understand image according to the textual information, and understand text according to the associated image information. So the information of one modality is conditionally fused with that of another modality for once or multiple times. Intuitively, there are inter-modality attention relations between image and text. However, existing state-of-the-art methods are weak to fuse multimodal features due to their neglect of inter-modality interactions.

To address the aforementioned challenges, we propose the Multimodal Co-Attention Networks (MCAN) for fake news detection by considering multimodal features. In our proposed model, we first extract spatial-domain features and frequency-domain features from image, as well as textual features from text. Then we develop a novel fusion approach with multiple co-attention layers to learn inter-modality relations, which fuses visual features first, and then the textual features. The fused representation obtained from the last co-attention layer is used for fake news detection.

The contributions of this paper can be summarized as follows: (1) We propose a novel end-to-end approach to detect fake news on social media only using the text and the attached image, without any extra information and auxiliary tasks. (2) The proposed MCAN model stacks multiple co-attention layers to fuse the multimodal features, which can learn inter-dependencies among them. (3) Our MCAN model is a general framework for fake news detection, and the components of MCAN are flexible. The sub-networks used to extract multimodal features can be replaced by different models. Moreover, the modular fusion process of MCAN allows our model to handle more modalities conveniently. (4) We evaluate MCAN on two large scale real-world datasets. The results demonstrate that our model outperforms the state-of-the-art models.

The rest of the paper is organized as follows: In Section 2, we summary previous related work on fake news detection. In Section 3, we detail our proposed model. The datasets, baselines, and experiment results are presented in Section 4. We conclude the study in Section 5.

2 Related Work

Following the previous work (Ruchansky et al., 2017; Shu et al., 2017), we specify that **fake news is the news that is intentionally fabricated and can be verified as false**. Existing methods for fake news detection can be divided into unimodal approaches and multimodal approaches.

2.1 Unimodal Fake News Detection.

Textual features are extracted from text content, including statistical features, such as the number of paragraphs in the text (Volkova et al., 2017), the percentage of negative words (Potthast et al., 2017; Bond et al., 2017), the number of punctuation and emojis (Castillo et al., 2011), and semantic features, such as writing styles (Chen et al., 2015) and language styles (Feng et al., 2012). However, these features are hand-designed, bringing bias and design difficulty. To address this problem, many studies use deep learning technologies, such as RNN (Ma et al., 2016), CNN (Yu et al., 2017), and GAN (Ma et al., 2019), to identify fake news. Their results show that deep learning methods perform better.

Visual features are important for news verification (Jin et al., 2016; Shu et al., 2017), such as clarity score (Jin et al., 2016), the number of images (Wu et al., 2015; Jin et al., 2016). However, these features are manually crafted and just learn simple patterns, hardly applying to real images. Qi et al. (2019) design a CNN-based model to capture image patterns, but their model only works in the case of large samples. So the applicable scope is very limited.

Social context features are born in the social connection between users and tweets, such as user profile and the number of posts. Liu et al. (2018) use user profiles on the news propagation path to determine the truth of the news. Some other works model propagation patterns as tree structures based on kernel methods (Wu et al., 2015; Ma et al., 2017). However, social context features are hand-crafted, incomplete, and unstructured.

The above work embodies the limitations of unimodal features in detecting fake news. In this paper,

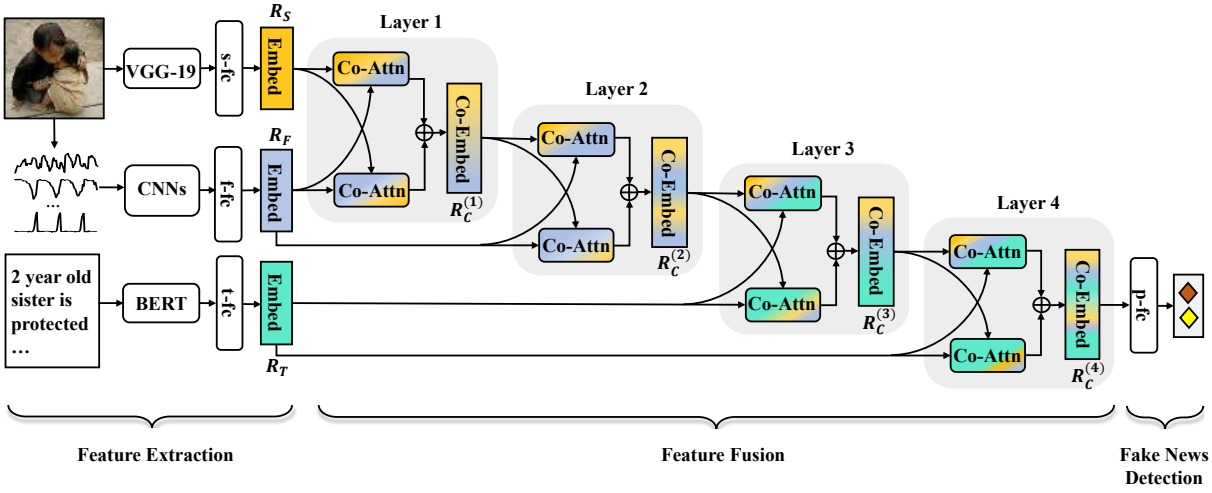


Figure 2: The architecture of our MCAN model.

we consider multiple modalities simultaneously when detecting fake news.

2.2 Multimodal Fake News Detection.

Recent works explore to fuse multimodal features. Jin et al. (2017) use local attention mechanism to fuse textual, visual, and social context features. Wang et al. (2018) learn event-invariant features by an aided adversarial network. Khattar et al. (2019) utilize autoencoders coupled with a detector to learn the shared representation of the text and the attached image. However, they ignore the characteristics of fake images at physical level (e.g., re-compression artifacts), and the fused features they learned lack correlations across multiple modalities.

To overcome the limitations of existing works, we propose MCAN to learn inter-dependencies among modalities. We extract spatial-domain and frequency-domain features of image, and textual features. Then we fuse them through a deep co-attention model inspired by a realistic scenario.

3 Methodology

3.1 Model Overview

Our model aims to learn multimodal fusion representations by considering dependencies across the modalities. As shown in Figure 2, the proposed model has three major procedures: feature extraction, feature fusion, and fake news detection.

Given news with text and image, we first utilize three different sub-models to extract features from spatial domain, frequency domain, and text. Then the multi-modality features are fused through a deep co-attention model, which consists of multiple co-attention layers. At last, the output of the co-attention model is used for judging the truth of the input news.

3.2 Feature Extraction

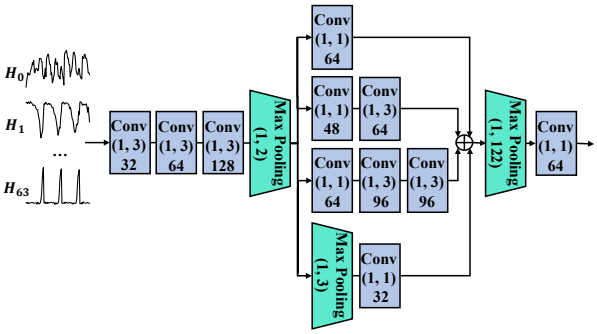


Figure 3: The detailed architecture of feature extraction in frequency domain.

Spatial-Domain Feature. To learn the semantic-level characteristics of the given image, we employ the VGG-19 network (Simonyan and Zisserman, 2014) to extract visual features from spatial domain. After the second of the last layer of VGG-19, we add a fully connected layer (denoted as "s-fc" in Figure 2) with ReLU activation function to generate a $d \times 1$ dimensional feature representation of the input image in spatial domain, which is denoted as $R_S \in \mathbb{R}^{d \times 1}$.

Frequency-Domain Feature. Fake-news images are often re-compressed images or tampered images that show periodicity in frequency domain (Qi et al., 2019), which can be easily captured by CNNs. Thus we design a CNN-based sub-network to extract features from frequency domain, as in Figure 3. The image is transformed from spatial

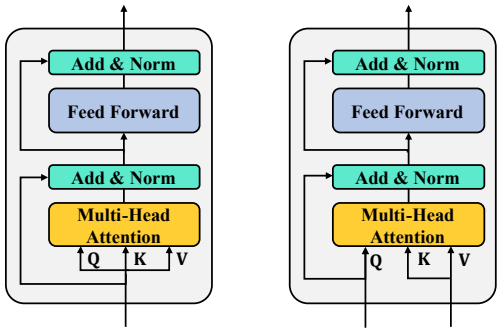
domain to frequency domain through discrete cosine transform (DCT) as in Qi et al. (2019). After that, we obtain 64 vectors $H_0, H_1, \ldots, H_{63}$, which are then sampled to the fixed size 250. For parallel computation, we pick 64 250-dimensional vectors into a matrix $H_F \in \mathbb{R}^{(64\times250)}$, which is fed to the CNN-based network later. The CNN-based subnetwork consists of a major network along with multi-branch networks. The earlier parts of the major network have three convolutional blocks and a max-pooling layer. The multi-branch networks are the same as architectures in Inception V3 (Szegedy et al., 2016). The last parts of the major network are a max-pooling layer followed by a convolutional block. Each convolutional block is comprised of a two-dimensional convolutional layer with batch normalization and ReLU activation function. After adding a fully connected layer with ReLU activation function (denoted as "f-fc" in Figure 2), we obtain the feature representation of the image in frequency domain $R_F \in \mathbb{R}^{d\times1}$.

Textual Feature. The text content of the tweet is a sequential list of words denoted as $T = [T_1, T_2, \ldots, T_n]$, where n is the number of words in a tweet, and each word $T_i \in T$ is tokenized by a pre-prepared vocabulary (Devlin et al., 2018). Recently, the BERT model (Devlin et al., 2018) which is pre-trained on a large language corpus, has been proven to perform very well in multiple natural language processing tasks. Thus we utilize BERT to obtain the aggregate sequence representation as textual features we desired. The textual feature is then resized to be a $d \times 1$ dimensional representation (denoted as R_T) by a fully connected layer with ReLU activation function.

3.3 Feature Fusion

Intuitively, people often look at the image first and then read the text when reading the news with image and text. This process may be repeated several times, continuously fusing image and text information. Therefore, we develop a novel fusion approach to simulate this process. Before presenting the fusing process, we first introduce its basic unit, the co-attention (CA) block. We achieve feature fusion by cascaded stacking multiple CA layers, which consists of two parallelly connected CA blocks.

Co-attention block. Co-attention block (Lu et al., 2019) is a variant of the standard multi-head self-attention (MSA) block (Vaswani et al., 2017),



(a) Self-attention block (b) Co-attention block

Figure 4: Illustration of the self-attention block and the co-attention block.

which can capture global dependencies of all positions in a sequence and is widely used in NLP and VQA tasks (Nguyen and Okatani, 2018; Gao et al., 2019). The MSA block showed in Figure 4(a) consists of a multi-headed self-attention function and a fully connected feed-forward network, both wrapped a residual connection followed by layer normalization. The input of MSA is first used to compute $(d \times 1)$-dimensional queries, keys, and values packed into matrixes Q, K, V, respectively. The similarity of the dot product between Q and K determines the attention distribution on the V. Multi-head attention function with m heads has m self-attention functions in parallel. For the i-th head, the inputs are transformed from Q, K, and V as follow:

$$Q_i = QW_i^Q,\ K_i = KW_i^K,\ V_i = VW_i^V \quad (1)$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{1\times d_h}$ are the projection matrices for the i-th head, $d_h = d/m$ is the dimensionality of the output feature of each head.

The calculation process of multi-head self-attention function can be presented as follows:

$$\begin{aligned} \mathrm{MA}(Q,K,V) &= \mathrm{h}\,W^O \\ \mathrm{h} &= \mathrm{h}_1 \oplus \mathrm{h}_2 \oplus \ldots \oplus \mathrm{h}_\mathrm{m} \\ \mathrm{h}_i &= \mathrm{A}(Q_i,K_i,V_i) = \mathrm{softmax}(\frac{Q_iK_i^T}{\sqrt{d_h}})V_i \end{aligned} \quad (2)$$

where $W^O \in \mathbb{R}^{md_h\times1}$, $\oplus$ denotes concatenation of vectors.

The fully connected feed-forward network consists of two linear transformations with a ReLU activation function in between,

$$\mathrm{FFN}(x) = \max(0, xW_1)W_2 \quad (3)$$

where the dimensionality of input and output is $d \times 1$, and the inner-layer dimensionality is d_{ff}.

The co-attention block (denoted as "Co-Attn") is extended from the MSA block, as shown in Figure 4(b). For a Co-Attn block, the queries are from one modality while keys and values are from another modality. Especially, the query matrice is used as a residual item after the multi-head attention sub-layer. The rest architectures are the same as MSA. The Co-Attn block produces an attention-pooled feature for one modality conditioned on another modality. If Q comes from text and k and V come from the attached image, the attention value calculated using Q and K can be used as a measure of the similarity between the text and image, and then weights the image. Just like humans, after reading the text, they will pay more attention to the areas in the image that are similar to the text. We believe that co-attention can simulate this process and learn inter-dependencies between different features.

Co-attention layer. We obtain a CA layer by connecting two Co-Attn blocks in parallel, as shown in Figure 2. Giving two Co-Attn blocks different features, the CA layer computes queries, keys, and values for each Co-Attn block as in a MSA block. Then the keys and values of one Co-Attn block are passed as input to another Co-Attn block. The outputs of two Co-Attn blocks are concatenated together and then fed into a fully connected layer to get the fused representation. The CA layer models dense interactions between input modalities by exchanging their information.

Multiple co-attention stacking. In order to fuse multimodal features deeply, we stack 4 CA layers in depth. The fusion process is progressive, and the output of each CA layer is one of the inputs of the next layer (see Figure 2). We first fuse spatial-domain representation R_S and frequency-domain representation R_F in first CA layer and obtain $R_C^{(1)}$. Then R_F are enhanced to fuse with $R_C^{(1)}$ in the second CA layer which outputs $R_C^{(2)}$. In the third and fourth layers, the inputs are the output of the previous layer and text representation R_T, and outputs are $R_C^{(3)}$ and $R_C^{(4)}$, respectively. The output vector of each CA layer is d-dimensional.The calculation processes are formulated as follows. Due to the page limit, we only show the calculation processes in the first CA layer and skip the repeated calculation details of other layers.

$$R_{C_{S\leftarrow F}} = R_S + \mathrm{MA}(R_S, R_F, R_F) \quad (4)$$

$$R_{C'_{S\leftarrow F}} = R_{C_{S\leftarrow F}} + \mathrm{FFN}(R_{C_{S\leftarrow F}})) \quad (5)$$

$$R_{C_{F\leftarrow S}} = R_F + \mathrm{MA}(R_F, R_S, R_S) \quad (6)$$

$$R_{C'_{F\leftarrow S}} = R_{C_{F\leftarrow S}} + \mathrm{FFN}(R_{C_{F\leftarrow S}}) \quad (7)$$

$$R_C^{(1)} = (R_{C'_{S\leftarrow F}} \oplus R_{C'_{F\leftarrow S}})W_C^{(1)} \quad (8)$$

where $R_{C'_{S\leftarrow F}} \in \mathbb{R}^d$ is the attention-pooled feature for spatial domain conditioned on frequency domain, $R_{C'_{F\leftarrow S}} \in \mathbb{R}^d$ is the attention-pooled feature for frequency domain conditioned on spatial domain, and $W_C^{(1)} \in \mathbb{R}^{2d\times d}$ is the projection matrice of the first CA layer. $R_C^{(1)}$ is transformed to be a $(d \times 1)$-dimensional representation before being input to the next CA layer. Specifically, the first and the third CA layers share parameters, and the second and the fourth CA layers share parameters.

3.4 Model Learning

We have obtained the multimodal feature representation $R_C^{(4)}$ fused features of text, spatial domain, and frequency domain. Let $f = R_C^{(4)}$, which is used to predict. The output of the proposed MCAN is the probability of a tweet being fake:

$$\hat{y} = \mathrm{softmax}\left(\max(0, fW_f)W_s\right) \quad (9)$$

where W_f is parameters of the fully connected layer, and W_s is parameters of the linear layer in the softmax layer. The loss function is devised to minimize the cross-entropy value:

$$L\left(\Theta\right) = -y\log\left(\hat{y}\right) - (1-y)\log\left(1-\hat{y}\right) \quad (10)$$

where y is the ground truth, with 1 representing fake news and 0 representing real news, and Θ denotes all learnable parameters in the proposed model.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed MCAN, we conduct experiments on two public real-world datasets, which are collected from Twitter and Weibo, respectively. The Twitter dataset was released for Verifying Multimedia Use task at MediaEval (Boididou et al., 2016). The Weibo dataset is collected by Jin et al. (2017). In the

	Twitter	Weibo
# of fake news	8199	4211
# of real news	6681	3639
# of images	512	7850

Table 1: Statistics of two datasets.

Weibo dataset, the real news is verified by an authoritative news agency in China, Xinhua News Agency. The fake news is verified by the official rumor debunking system of Weibo. The tweets in each dataset contain texts, attached images/videos, and social context information. In this work, we focus on text and image information. So we remove the tweets with videos and the tweets without texts or images. In Twitter dataset, 512 images are shared by the remaining data. When preprocessing the Weibo dataset, the steps we used are similar to that in the work (Jin et al., 2017). We keep the same data split scheme as the benchmark on these two datasets. The detailed statistics of the two datasets are listed in Table 1.

4.2 Experimental Settings

The max length of the text is 25 on Twitter and 160 on Weibo. The hidden size of "s-fc", "f-fc" and "t-fc" are 256. We set d=256, m = 4, and d_{ff} = 512. The hidden size of "p-fc" is 35. The parameters of VGG-19 and BERT are frozen when training on Twitter dataset due to overfitting, but not on Weibo dataset. The BERT model used on Twitter dataset is multilingual cased BERT-based model and the one used on Weibo dataset is Chinese BERT-based model. Our proposed model is trained for 100 epochs with early stopping. We use Adam (Kingma and Ba, 2014) and AdaBelief (Zhuang et al., 2020) as optimizers on Twitter and Weibo datasets, respectively, to seek the optimal parameters of our model. The optimal hyperparameters of our model are determined by grid searching, and the selection criterion is accuracy. The hyperparameters of baselines are the same as those in respective studies.

4.3 Baselines

To validate the effectiveness of MCAN, we choose two categories of baseline models: unimodal models and multimodal models, which are listed as follows: (1) **Text**: a BERT model coupled with the decision network in MCAN, using textual information. (2) **Spatial**: a model consists of a VGG-19 model and the decision network of MCAN, utilizing image information in spatial domain. (3) **Freq**: proposed MCAN only has the part of dealing with frequency-domain features. (4) **VQA** (Antol et al., 2015): a model aims to answer questions according to the given images. For fair comparisons, we use a one-layer LSTM. (5) **NeuralTalk** (Vinyals et al., 2014): a deep recurrent framework for image caption. The joint representation of image and text is obtained by averaging the output of RNN at each timestep. (6) **att-RNN** (Jin et al., 2017): att-RNN utilizes local attention to fuse textual, visual, and social context features. For a fair comparison, we remove the part dealing with social context information. (7) **EANN** (Wang et al., 2018): A neural network based on the adversarial idea to remove the event-specific features. In EANN, event identification is an auxiliary task, and event labels are not in original datasets. For a fair comparison, we removed the event discriminator. (8) **MVAE** (Khattar et al., 2019): MVAE learns shared representations of text and image using a variational autoencoder coupled with a binary classifier. We use the same model as in the original work (Khattar et al., 2019). (9) **MCAN-A**: MCAN without the part of fusing multimodal features. Spatial-domain features, frequency-domain features, and textual features are simply concatenated for prediction.

4.4 Performance Comparison

Table 2 shows the results of baselines and our proposed model on two datasets. We can observe that the proposed MCAN outperforms all the baselines over all metrics across two datasets.

There are many similar trends on the two datasets. MCAN-A performs better than unimodal models, which indicates that adding features usually improves model performance, but it is not always positively correlated. For example, Text on Weibo dataset is better than MCAN-A. After adding the process of multimodal fusion, our proposed MCAN beats MCAN-A and other multimodal models, which embodies our proposed feature fusion method is indeed better than the simple concatenation method.

There are also some differences on the two datasets. The performance of Text (BERT) and Spatial (VGG-19) on Weibo dataset is much better than that on Twitter dataset. The reason is related to the dataset itself. On Weibo dataset, the average length of a tweet is about 10 times of that of a tweet

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Text	0.633	0.656	0.762	0.705	0.587	0.459	0.515
	Spatial	0.671	0.841	0.527	0.648	0.574	0.864	0.69
	Freq	0.665	0.733	0.656	0.692	0.592	0.677	0.631
	VQA	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.81	0.498	0.617	0.584	0.759	0.66
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.73
	MCAN-A	0.737	0.840	0.671	0.746	0.65	0.827	0.727
	MCAN	**0.809**	**0.889**	**0.765**	**0.822**	**0.732**	**0.871**	**0.795**
Weibo	Text	0.876	0.885	0.871	0.878	0.865	0.878	0.871
	Spatial	0.857	0.85	0.877	0.863	0.863	0.834	0.848
	Freq	0.717	0.728	0.724	0.726	0.706	0.710	0.708
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	NeuralTalk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.72	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
	MCAN-A	0.869	0.868	0.879	0.874	0.869	0.857	0.863
	MCAN	**0.899**	**0.913**	**0.889**	**0.901**	**0.884**	**0.909**	**0.897**

Table 2: The results of different methods on two datasets

on Twitter dataset, which probably makes BERT perform better on Weibo dataset. Moreover, more than 70% of tweets on Twitter dataset are related to a single event. Thus, the training samples of BERT and VGG-19 are too similar, resulting in poor performance of model generalization. This is the reason why we fine-tuned BERT and VGG-19 on Weibo dataset but not on Twitter dataset. They are easy to overfit on Twitter dataset. But Weibo dataset has no such imbalanced issue.

On Weibo dataset, the accuracy of fine-tuned BERT and VGG-19 all exceed 85%. In this case, our proposed MCAN further improves the accuracy to close to 90% with the help of cascaded way of stacking CA layers. Comparing with the situation on Twitter dataset, we can find that our model performs better in the face of weak unimodal features. In our MCAN model, the representation ability of features can be greatly improved by effectively fusing other features.

4.5 Ablation Analysis

Quantitative Analysis. To evaluate the effectiveness of each component of the proposed MCAN, we remove each one from the entire model for comparison. "ALL" denotes the entire model MCAN with all components, including spatial-domain representation (S), textual representation (T), frequency-domain representation (F), and co-attention layers (A). After removing each one of them, we obtain the sub-models "-S", "-T", "-F" and "-A", respectively. "-F-A" denotes the reduced MCAN without both frequency-domain representation and co-attention layers. The results are exhibited in Figure 5.

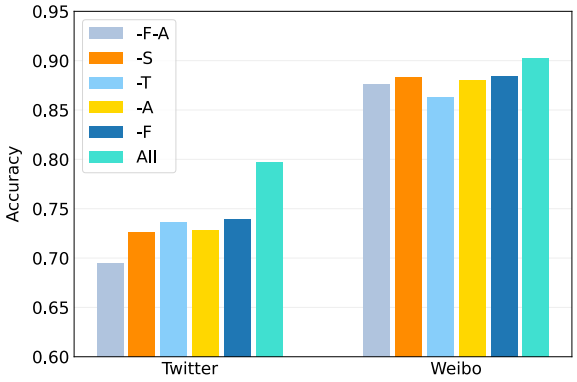
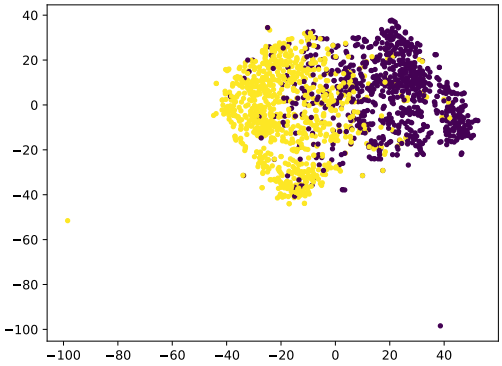


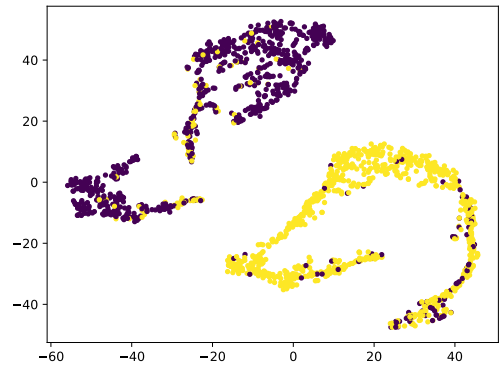
Figure 5: MCAN ablation analysis in Accuracy.

We can see that every component plays a significant role in improving the performance of MCAN. MCAN beats MCAN-F, which reveals that the frequency domain information is indeed helpful to detect fake news. On Twitter dataset, the contribution of textual representations to the entire model is less than that of visual representations, while the situation on Weibo dataset is opposite. This is still due to the imbalanced issue and the less average length of a tweet on Twitter dataset, which decrease the performance of the textual represen-

tation. Besides, on Weibo dataset, removing one or two components, the performance of MCAN does not drop significantly as on Twitter dataset. This benefits from balanced data distribution and the stability of fine-tuned BERT and VGG-19, as mentioned in Section 4.4.



(a) MCAN-A



(b) MCAN

Figure 6: Visualizations of learned latent feature representations.

Qualitative Analysis. To illustrate the effectiveness of co-attention layers in MCAN, we qualitatively visualize the joint representation of three modalities learned by MCAN-A and the fused representation $R_C^{(4)}$ learned by MCAN on Weibo testing set with t-SNE (Maaten and Hinton, 2008), as shown in Figure 6. The label of each tweet is real or fake.

From Figure 6, we can observe that the separability of the feature representation learned by MCAN is much better than its reduced model MCAN-A. MCAN-A can learn discriminable features, but many features are still easily misclassified, showing in Figure 6(a). On the contrary, the features learned by MCAN are more discriminable with a more significant segregated area between two types of samples, as exhibited in Figure 6(b). This is attributed to the cascaded way of stacking co-attention layers in MCAN, which fuses the characteristics of multiple modalities deeply and boosts to distinguish fake news and real news.

From the above phenomena, we can conclude that the proposed method MCAN learns better and more distinctive feature representations with the co-attention layers, thus achieving better performance.

4.6 Case Studies

To further illustrate the importance of multimodal features for fake news detection, we compare the results reported by MCAN and unimodal models (Text and Spatial) and exhibit some fake news correctly captured by MCAN but missed by unimodal models.

Before washed away by flood, an Indian man calmly gave the last gesture to a photographer.

A group of dolphins brought a dog that fell into a canal to safe area.

Figure 7: Some fake news detected by MCAN but missed by Text on the Weibo dataset.

Figure 7 shows two top-confident tweets successfully detected by MCAN but missed by text-only MCAN. The textual contents of the two examples can provide little evidence that it is fake news. However, the two attached images seem forged pictures.

The water mantis lives in sewers. Its head has two to three times the poison of pufferfish and has no antidote.

Several urban management officers are frantically plundering street-side property worth more than 100 million yuan.

Figure 8: Some fake news detected by MCAN but missed by Spatial on the Weibo dataset.

In Figure 8, the two examples are detected by MCAN but missed by Spatial. The attached images in two examples look normal. However, the words in the tweet seem exaggerated and unbelievable. It is challenging for spatial-domain-only MCAN to detect, but with multimodal features, our MCAN model identifies them correctly.

These comparative cases prove that when a single-modal model, whether a text-based model or an image-based model, cannot correctly distinguish fake news, the proposed MCAN using multimodal features can give high confidence.

5 Conclusions

In this work, we propose a novel Multimodal Co-Attention Networks (MCAN) to tackle the challenge of fusing multimodal (textual and visual) features for fake news detection. We utilize three different sub-networks to extract features from text, spatial domain, and frequency domain, respectively. Then the three features are deeply fused by stacking co-attention layers, which is inspired by human behavior. When people read news with image, image and text are read once or multiple times, and continuously fused in brain. Experiments on two public benchmark datasets for fake news detection validate the effectiveness of MCAN, and the results show that MCAN outperforms the current state-of-the-art methods. In the future, we plan to extend the co-attention based fusion approach in MCAN to other fake news research, such as fake news diffusion.

Acknowledgments

This research was supported by National Research and Development Program of China (No.2017YFB1010004).

References

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, and D. Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Los Alamitos, CA, USA. IEEE Computer Society.

C. Boididou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, and Y. Kompatsiaris. 2016. Verifying multimedia use at mediaeval 2016. In *MediaEval 2016 Workshop*.

Gary D Bond, Rebecka D Holman, Jamie-Ann L Eggert, Lassiter F Speller, Olivia N Garcia, Sasha C Mejia, Kohlby W Mcinnes, Eleny C Ceniceros, and Rebecca Rustige. 2017. 'lyin'ted','crooked hillary', and 'deceptive donald': Language of lies in the 2016 us presidential debates. *Applied Cognitive Psychology*, 31(6):668–677.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, pages 3049–3055.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 518–527. IEEE.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. 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*, pages 849–857.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. pages 540–547.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A convolutional approach for misinformation identification.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. pages 1395–1405.

Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James S. Duncan. 2020. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients.