Frequency Spectrum Is More Effective for Multimodal Representation and Fusion: A Multimodal Spectrum Rumor Detector

An Lao[1], Qi Zhang[2,4], Chongyang Shi[1]*, Longbing Cao[3], Kun Yi[1], Liang Hu[2,4], Duoqian Miao[2]

[1]Beijing Institute of Technology
[2]Tongji University
[3]Macquarie University
[4]DeepBlue Academy of Sciences

{an.lao, cy_shi,yikun}@bit.edu.cn, {zhangqi_cs, lianghu, dqmiao}@tongji.edu.cn, longbing.cao@mq.edu.au

Abstract

Multimodal content, such as mixing text with images, presents significant challenges to rumor detection in social media. Existing multimodal rumor detection has focused on mixing tokens among spatial and sequential locations for uni-modal representation or fusing clues of rumor veracity across modalities. However, they suffer from less discriminative uni-modal representation and are vulnerable to intricate location dependencies in the time-consuming fusion of spatial and sequential tokens. This work makes the first attempt at multimodal rumor detection in the frequency domain, which efficiently transforms spatial features into the frequency spectrum and obtains highly discriminative spectrum features for multimodal representation and fusion. A novel **F**requency **S**pectrum **R**epresentation and f**U**sion network (FSRU) with dual contrastive learning reveals the frequency spectrum is more effective for multimodal representation and fusion, extracting the informative components for rumor detection. FSRU involves three novel mechanisms: utilizing the Fourier transform to convert features in the spatial domain to the frequency domain, the unimodal spectrum compression, and the cross-modal spectrum co-selection module in the frequency domain. Substantial experiments show that FSRU achieves satisfactory multimodal rumor detection performance.

Introduction

With the rapid development of social media in various aspects of our lives, the prevalence of content from multiple sources and in diverse formats has significantly increased. A prime example is the combination of text of varying lengths accompanied by images. However, along with this proliferation of multimodal media, a more sophisticated and concerning issue has arisen: multimodal rumors. Multimodal rumors refer to disseminating misinformation or false information through social media platforms, incorporating multiple modes of communication such as text and images. These rumors often defy logical reasoning and lack credibility. Research reveals that rumors are shared more extensively on Facebook than on mainstream news (Willmore 2016). As a result, it has become imperative to detect and mitigate multimodal rumors to effectively manage the associated risks and ensure compliance with social media norms and guidelines (Allcott and Gentzkow 2017; Zhang et al. 2023).

Recent studies of multimodal rumor detection primarily focus on two key aspects: learning spatial and sequential dependencies in uni-modality and fusing evidence of rumor veracity across different modalities (Chen et al. 2022; Zheng et al. 2022; Singhal et al. 2022). 1) To obtain informative uni-modal representation, researchers have employed various neural models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers to perform token mixing over spatial locations of images or sequential positions of text. However, these methods suffer from less discriminative unimodal representation, hindering subsequent fine-grained cross-modal fusion. 2) Existing approaches often apply contrastive learning (Ying et al. 2023) or co-attention mechanisms (Qian et al. 2021) to achieve multimodal alignment or fusion for detecting rumor across modalities. However, they may either overlook the interpretable fine-grained fusion or encounter intricate location dependencies in fusing spatial and sequential tokens. Moreover, current approaches for fine-grained fusion, such as co-attention mechanisms, often exhibit quadratic time complexity (Rao et al. 2021). These issues collectively undermine the accuracy and efficiency of multimodal rumor detection models, highlighting the need for further advancements in this field.

To address the issues, we make the first attempt from a new paradigm and architecture in this work: multimodal spectrum rumor detection. We contend that the frequency spectrum offers a more effective means of representing and fusing multimodal data. Inspired by signal processing theories (Mateos et al. 2019), we can utilize Fourier transforms to transform sequential (text) or spatial (images) data to the frequency domain. The Fourier transform often generates *a sparse frequency spectrum* with a significant portion of frequency components approaching zero (shown in Figure 1). This characteristic facilitates obtaining discriminative uni-modal representation and emphasizing (suppressing) veracity-relevant (irrelevant) features for detection. In addition, the frequency spectrum provides *a global view* (Rao et al. 2021), allowing each spectrum component to attend to all features in the spatial domain. Unlike the position-based alignment in co-attention mechanisms (Zheng et al.
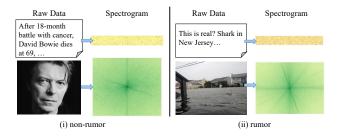
Figure 1: Two examples on Twitter visualize the raw data and its spectrograms. It shows the spectrum has centrally concentrated components and discriminative patterns.

2022), the spectrum exhibits global patterns (see Figure 1), allowing a more comprehensive sense of intricate location dependencies within/across modalities between rumors and non-rumors. Moreover, point-wise multiplication in the frequency domain is equivalent to self-attention in the spatial domain, avoiding quadratic time complexity (Appendix A).

Accordingly, we propose an architecturally simple and computationally efficient multimodal spectrum rumor detector: a Frequency Spectrum Representation and fUsion network (FSRU) with dual contrastive learning. FSRU comprises three key components: text and image embedding, multimodal frequency spectrum representation and fusion module, and detection with distribution similarity. Especially, the frequency spectrum representation and fusion module includes four core operations: we introduce 1) discrete Fourier transform (DFT) to convert features in the spatial domain to the frequency domain; 2) unimodal spectrum compression to compress frequency domain features; 3) cross-modal spectrum co-selection to select spectrum components; and 4) inverse DFT (IDFT) to reverse frequency domain features to the spatial domain. By utilizing filter banks in the frequency domain, unimodal spectrum compression generates spectral compressed representations to reveal potential features within each modality and portray distinct feature patterns. Cross-modal spectrum co-selection makes use of complementary dependencies between modalities to select informative spectrum components that are beneficial in identifying rumors. Subsequently, we devise a fusion module that leverages the similarity of feature distributions to generate a cohesive multimodal representation and introduce dual contrastive learning to enhance multimodal learning. We conduct experiments on two real-world datasets to evaluate our proposed approach, FSRU. The results demonstrate that FSRU yields favorable outcomes across different evaluation metrics and aspects.

Our contributions are twofold:

- An architecturally simple and computationally efficient novel method **F**requency **S**pectrum **R**epresentation and f**U**sion network (FSRU) with dual contrastive learning is proposed for multimodal rumor detection. Unlike existing approaches that primarily focus on features in the spatial/sequential domain, FSRU aims to capture discriminative unimodal features and fuse cross-modal evidence of rumor veracity in the frequency domain. This

architecturally simple approach offers a fresh perspective on multimodal rumor detection.

- A frequency spectrum representation and fusion module is proposed to extract rumor evidence that is concealed in the frequency components from both unimodal and cross-modal perspectives. The unimodal spectrum compression explores clearer patterns in text and image representations. The cross-modal spectrum co-selection guides retaining relevant frequency components while fusing multimodal spectrum features, effectively reducing the impact of irrelevant frequency components.

Related Work

Multimodal Rumor Detection

Previous work attempts to solve multimodal rumor detection by concatenating text and image features (Wang et al. 2018; Cui, Wang, and Lee 2019; Singhal et al. 2019; Zhang et al. 2020). They concatenate multimodal features from the spatial dimension without considering modal interactions. To address this deficiency, MFN (Chen et al. 2021) employs a self-attentive fusion module to capture the relationships between text and image. CAFE (Chen et al. 2022) introduces cross-modal alignment and ambiguity learning to learn cross-modal correlations while integrating multimodal features. Hidden state contextual information complements the modal representation during the feature representation phase. Sun et al. (Sun et al. 2021) design a modality-shared embedding and introduce external knowledge to assist with rumor detection. Recently, attention-based functions have been popularly involved in multimodal rumor detection. MFAN (Zheng et al. 2022) enhances the model representation by extracting mutual information between modalities through cross-modal co-attention mechanisms. To improve the multimodal learning capability, HMCAN (Qian et al. 2021) adopts a Transformer-based contextual attention network to extract multimodal contextual complementary information. BMR (Ying et al. 2023) proposes the Improved Multi-gate Mixture-of-Expert networks to learn information from unimodal and multimodal features through single-view prediction and cross-modal consistency learning.

Fourier Transform in Deep Learning

Fourier transform plays a vital role in the area of digital signal processing. It has been introduced to deep learning for enhanced learning performance (Ehrlich and Davis 2019; Chi, Jiang, and Mu 2020; Li et al. 2020; Yang and Soatto 2020; Yi et al. 2023c,a). GFNet (Rao et al. 2021) utilizes fast Fourier transform to convert images to the frequency domain and exchange global information between learnable filters. As a continuous global convolution independent of input resolution, Guibas et al. (Guibas et al. 2021) design the adaptive Fourier neural operator frame token mixing. Xu et al. (Xu et al. 2020) devise a learning-based frequency selection method to identify trivial frequency components and improve the accuracy of classifying images. On text classification, Lee-Thorp et al. (Lee-Thorp et al. 2022) use the Fourier transform as a text token mixing mechanism. Furthermore, the Fourier transform is also applied to forecast

time series (Cao et al. 2020; Lange, Brunton, and Kutz 2021; Koç and Koç 2022; Yang and Hong 2022). To increase the accuracy of multivariate time-series forecasting, Cao et al. (Cao et al. 2020) propose a spectral temporal graph neural network (StemGNN), which mines the correlations and time dependencies between sequences in the spectral domain. Yang et al. (Yang and Hong 2022) propose bilinear temporal spectral fusion (BTSF), which updates the feature representation in a fused manner by explicitly encoding time-frequency pairs and using two aggregation modules: spectrum-to-time and time-to-spectrum.

Our work is inspired by (Rao et al. 2021; Xu et al. 2020; Yi et al. 2023b) but differs from them. To our knowledge, there are no existing techniques for multimodal rumor detection that employ the same architecture for frequency domain characterization as our approach. Our approach differs from other spatial domain techniques in that we not only convert the original features into the frequency domain but also perform a series of complex-valued computation operations in the frequency domain.

Problem Definition

We formulate multimodal rumor detection as a binary classification task, where multimodal a refers to text and image modalities, denoted as $a \in \{t, v\}$. Given a multimodal rumor dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, each sample is denoted as (x, y), and x can be represented by $x = \{x^t, x^v\}$, where x^t stands for text and x^v for image. $y \in \{0, 1\}$ is the rumor veracity label corresponding to sample x, $y = 1$ indicates that the sample is a rumor, while $y = 0$ indicates that the sample is true. This work aims to incorporate text and image features to predict the rumor label $\hat{y} \in \{0, 1\}$.

Methodology

We propose a Frequency Spectrum Representation and fUsion network (FSRU) with dual contrastive learning to tackle the problem of multimodal rumor detection. As illustrated in Figure 2, FSRU comprises three components: 1) *text and image embedding* obtains textual and visual unimodal embeddings for social media posts through two embedding modules, respectively. 2) *frequency spectrum representation and fusion module* explores unimodal spectrum information and cross-modal spectrum interactions. 3) *detection with distribution similarity* performs the final detection after obtaining a multimodal representation by capturing the complementary semantic relationships between unimodality. Next, we explain each component in detail.

Text and Image Embedding

Given a rumor sample $x = \{x^t, x^v\}$, we first embed its raw text and image, respectively. Regarding the text sequence $x^t = [w_1, w_2, ..., w_m]$ (m is the number of words), we simultaneously employ word embedding and positional embedding to encode each word, denoted by:

$$\mathbf{w}_i = \text{WE}(w_i) + \text{PE}^t(w_i) \qquad (1)$$

where $\text{WE}(\cdot)$ is the word embedding and $\text{PE}^t(\cdot)$ is the position embedding for text sequence. Accordingly, we obtain

the text embedding $\mathbf{x}^t = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m]$. Regarding images, we divide each image into $h \times w$ non-overlapping patches $x^v = [p_1, p_2, ..., p_n]$ ($n = h \times w$) and adopt CNN (LeCun, Bengio et al. 1995) to generate meaningful representations:

$$\mathbf{p}_i = \text{CNN}(\text{PE}^v(p_i)) \qquad (2)$$

where $\text{PE}^v(\cdot)$ is the patch embedding for the image. We obtain the image embedding $\mathbf{x}^v = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n]$.

Frequency Spectrum Representation and Fusion

The frequency spectrum representation and fusion module losslessly transform spatial domain features into the frequency domain, obtaining discriminative spectrum features for each modality. The frequency spectrum gives text and image representations a complete view of spatial features and facilitates obtaining informative components and eliminating irrelevant components from a global view.

Spectrum Representation We first transform the spatial features into spectrum features using Discrete Fourier transform (DFT). The spectrum of text features can be obtained as follows:

$$\mathbf{X}^t[k] = \mathcal{F}_{seq}(\mathbf{x}^t[i]) = \sum_{i=0}^{m-1} \mathbf{x}^t[i]e^{-j(2\pi/m)ki} \qquad (3)$$

where $\mathbf{X}^t \in \mathbb{C}^{m \times d}$ is a complex tensor, $\mathbf{X}^t[k]$ is the spectrum of $\mathbf{x}^t[i]$ at the frequency $2\pi k/m$, $\mathcal{F}_{seq}(\cdot)$ is the 1D DFT along the sequence dimension, and j is the imaginary unit. The spectrum of image embedding can be obtained:

$$\mathbf{X}^v[k] = \mathcal{F}_{pat}(\mathbf{x}^v[i]) = \sum_{i=0}^{n-1} \mathbf{x}^v[i]e^{-j(2\pi/n)ki} \qquad (4)$$

where $\mathbf{X}^v \in \mathbb{C}^{n \times d}$ is a complex tensor, $\mathcal{F}_{pat}(\cdot)$ denotes the 1D DFT along the patch dimension. Self-attention computes the spatial dependencies in a quadratic time complexity, while DFT can be efficiently implemented via a fast Fourier transform in logarithmic time complexity. Refer to Appendix C for a more detailed comparison.

Unimodal Spectrum Compression (USC) Spatial features are effectively consolidated within each frequency element, enabling the extraction of informative features from both text and images through the point-wise product in the frequency domain. We introduce a filter bank for each modality $\mathbf{X}^a, a \in \{t, v\}$ to compress the spectrum and obtain the significant features associated with rumors. We use $\mathbf{K}^a = [\mathbf{k}_1^a, \mathbf{k}_2^a, ..., \mathbf{k}_k^a]$ to represent the filter bank, where k is the number of filters in the filter bank:

$$\hat{\mathbf{X}}^a = \sum_{i=1}^{k} \frac{1}{l}|\mathbf{X}^a|^2 \odot \mathbf{k}_i^a cos(\frac{(2i-1)\pi}{2k}), a \in \{t, v\} \quad (5)$$

where $\odot$ is the element-wise multiplication, $|\mathbf{X}^a|^2$ is the power spectrum of $\mathbf{X}^a$, l is the length of $\mathbf{X}^a$. The $|\mathbf{X}^a|^2$ operation smooths the spectrum, highlighting the main components of the spectrum from an intra-modal perspective. It also facilitates the subsequent learning of unimodal compression. $cos((2i-1)\pi/2k)$ compacts better energy and can
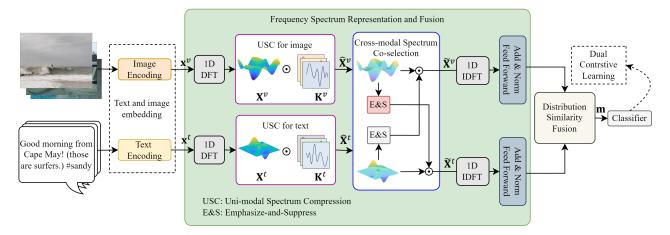
Figure 2: The architecture of our proposed Frequency Spectrum Representation and fUsion network (FSRU) for multimodal rumor detection. FSRU comprises three main components: a text and image embedding module, a frequency spectrum representation and fusion module, and a classification with distribution similarity.

aggregate the more important information in the rumor features. Its combination of application with the filter bank $\mathbf{K}^a$ allows for efficient frequency domain feature compression.

Cross-modal Spectrum Co-Selection (CSC) Based on the postulation that certain spectrum components have limited contributions to rumor detection, we propose an emphasize and suppress (E&S) module, which aims to enhance informative components and suppress irrelevant components within each modality by co-attending to the unimodal spectrum. We first perform average pooling over the compressed spectrum $\hat{\mathbf{X}}^a, a \in \{t, v\}$, subsequently applying convolution to obtain the representation of the rumor visual/text clues. Consequently, we can derive two selection filters, one from the visual spectrum and another from the text spectrum. The filters serve the purpose of co-selecting informative features from each other. We perform cross-modal spectrum co-selection by multiplying the two filters with the corresponding unimodal spectrum in a staggered manner:

$$\tilde{\mathbf{X}}^t = \hat{\mathbf{X}}^t \odot \text{Conv}(\text{Avg}(\hat{\mathbf{X}}^v \odot \boldsymbol{\Theta}^v)) \quad (6)$$

$$\tilde{\mathbf{X}}^v = \hat{\mathbf{X}}^v \odot \text{Conv}(\text{Avg}(\hat{\mathbf{X}}^t \odot \boldsymbol{\Theta}^t)) \quad (7)$$

where $\odot$ is the element-wise multiplication, $\boldsymbol{\Theta}^a$ denotes the trainable parameters with the same dimension as $\hat{\mathbf{X}}^a$, $Conv(\cdot)$ is an 1×1 convolutional layer, and $Avg(\cdot)$ is the average pooling function. The convolutional layer and $\boldsymbol{\Theta}^a$ facilitate learning how to emphasize informative components and suppress irrelevant components for multimodal fusion.

Finally, we employ inverse discrete Fourier transform (IDFT, $\mathcal{F}_{seq}^{-1}$ and $\mathcal{F}_{pat}^{-1}$) to convert the spectral representations of text and image back into the spatial domain:

$$\mathbf{x}^t \leftarrow \mathcal{F}_{seq}^{-1}(\tilde{\mathbf{X}}^t) \quad (8)$$

$$\mathbf{x}^v \leftarrow \mathcal{F}_{pat}^{-1}(\tilde{\mathbf{X}}^v) \quad (9)$$

The fine-grained cross-modal spectrum co-selection facilitates the common analysis of spectral components in text

and images during the inference process and guarantees the fusion of multimodal rumor features, which allows the retention of the informative components more properly.

Rumor Detection with Contrastive Learning

Contrastive Learning Objectives To promote multimodal learning in training, we introduce a dual contrastive learning module, consisting of two parts: 1) fully-supervised intra-modal contrastive learning based on rumor veracity labels $\mathcal{L}_{full}$, and 2) self-supervised inter-modal contrastive learning based on multimodal spatial semantics $\mathcal{L}_{self}$.

In a mini-batch $\mathcal{B}$, we divide samples according to the rumor veracity label into R_0, R_1. For the anchor sample $r_i \in R_1$, the positive pair can be denoted as (r_i, r_j), where $r_j \in R_1, j \neq i$. The samples in R_0 are regarded as negative examples. As such, we follow (Lin et al. 2022) to define the pairwise objective function with anchor sample and positive or negative samples $\mathcal{L}_1(\mathbf{x}^a, \mathbf{x}^a), a \in \{t, v\}$. The final fully-supervised intra-modal contrastive loss is as follows:

$$\mathcal{L}_{full} = \sum_{\mathcal{M}} [\sum_{r_i \in R_1} \frac{1}{|R_1|} \sum_{j, r_j \in R_1, j \neq i} \mathcal{L}_1(\mathbf{x}_i^a, \mathbf{x}_j^a) + \\ \sum_{r_k \in R_0} \frac{1}{|R_0|} \sum_{l, r_l \in R_1, l \neq k} \mathcal{L}_1(\mathbf{x}_k^a, \mathbf{x}_l^a)] \quad (10)$$

where $|\cdot|$ denotes the number of corresponding samples.

For self-supervised inter-modal contrastive loss, we consider the text and associated image of the given anchor sample r_i to be a positive sample, while the other pairs are considered negative samples. We use the InfoNCE loss (He et al. 2020) to optimize the image and text features, denoted as $\mathcal{L}_2(\mathbf{x}^t, \mathbf{x}^v)$ and $\mathcal{L}_2(\mathbf{x}^v, \mathbf{x}^t)$. And the self-supervised inter-modal contrastive loss is as follows:

$$\mathcal{L}_{self} = \frac{1}{2|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} [\mathcal{L}_2(\mathbf{x}_i^t, \mathbf{x}_i^v) + \mathcal{L}_2(\mathbf{x}_i^v, \mathbf{x}_i^t)] \quad (11)$$

where $|\mathcal{B}|$ denotes the number of samples in mini-batch $\mathcal{B}$.

Detection Based on Distribution Similarity After obtaining the improved text and image representations, we measure the Jensen-Shannon (JS) divergence between the two features to learn the distribution similarity, which is subsequently utilized to control the final multimodal rumor representation output. Since it is difficult to infer the posterior probability p from the given data sample, we generate an approximation of its distribution q. Specifically, the posterior probability of unimodal can be denoted separately as $q(z^t|x^t)$ and $q(z^v|x^v)$. The divergence of different modal distributions in $\mathbf{x}^a$ can then be measured as follows:

$$\gamma = \text{JS}(q(z^t|x^t)||q(z^v|x^v)) \tag{12}$$

where $JS(\cdot)$ denotes the JS divergence, and the similarity score γ is computed by the JS divergence. Accordingly, we can calculate the integrated multimodal representation and apply a fully connected layer FC to predict the label $\hat{y}$:

$$\mathbf{m} = (1 - \gamma)(\mathbf{W}^t\mathbf{x}_t + \mathbf{W}^v\mathbf{x}_v) + \gamma\mathbf{x}_t + \gamma\mathbf{x}_v \tag{13}$$

$$\hat{y} = \text{Softmax}(\text{FC}(\mathbf{m})) \tag{14}$$

where $\mathbf{W}^t$ and $\mathbf{W}^v$ are trainable parameters, and γ is a hyperparameter to adaptively weigh cross-modal features.

Taking rumor detection as a binary classification task, we then apply the cross-entropy loss as the detection objective:

$$\mathcal{L}_{cls} = -\mathbb{E}_{y\sim\hat{Y}}[ylog(\hat{y}) + (1 - y)log(1 - \hat{y})] \tag{15}$$

Finally, the final loss can be written as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{full} + \beta\mathcal{L}_{self} \tag{16}$$

with hyperparameters α, β to balance different objectives.

Experiments

In this section, we evaluate the effectiveness of our proposed model [1] on two real-world datasets.

Experimental Setup

Datasets To facilitate comparison with the baselines, we evaluate the proposed FSRU on two publicly available multimodal datasets: Twitter (Boididou et al. 2014) and Weibo (Jin et al. 2017). We comprehensively describe each dataset in Appendix B.1.

Baselines We compare our FSRU to recent baseline models: att-RNN (Jin et al. 2017), EANN (Wang et al. 2018), MVAE (Khattar et al. 2019), SpotFake (Singhal et al. 2019), HCMAN (Qian et al. 2021), CAFE (Chen et al. 2022), BMR (Ying et al. 2023), and LogicDM (Liu, Wang, and Li 2023). We comprehensively describe each baseline in Appendix B.2 and explain the rationale behind selecting these specific baselines.

Settings We implemented our algorithms using PyTorch 1.12 and conducted all experiments on a single NVIDIA RTX 3080 Ti GPU. The loss function is optimized using the Adam algorithm (Kingma and Ba 2015). The evaluation metrics include Accuracy, Precision, Recall, and F1 score.

[1]https://github.com/dm4m/FSRU

To ensure fairness, we employ five-fold cross-validation for the experiments. We utilize publicly available Word2Vec (Mikolov et al. 2013) to obtain the word embeddings. Images are resized into 224×224. The maximum sequence length is set to 50 for Weibo and 32 for Twitter. The dimension of text and image embedding is set to 256. The model is trained for 50 epochs with a batch size of 64. For Weibo, the initial learning rate is set to 1e-2, while for Twitter, it is set to 1e-5. When selecting hyper-parameters α and β, we consider values from the set $\{0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Ultimately, we set α and β to 0.2 for both datasets. The number of filters in unimodal spectrum compression denoted as k is chosen from the set $\{1, 2, 4, 8\}$, and the final value selected for the results is $k = 2$. To efficiently implement the DFT and IDFT, we utilized the Fast Fourier Transform (FFT) and inverse FFT. The code and implementation details can be found in the supplementary materials.

Results and Analysis

The performance comparison between FSRU and eight other baselines on the two datasets is presented in Table 1. We further investigate the complexity of FSRU in terms of FLOPs and parameter volumes, compared with state-of-the-art methods. The results are shown in Appendix C.

Att-RNN, EANN, and MVAE overlook the deep semantic relationships and interactions among features, leading to limitations in their detection accuracy. SpotFake leverages pre-trained models to extract text and image features, demonstrating strong performance in classifying rumors but relatively weaker performance in classifying non-rumors. The Transformer is utilized as a feature encoder in HMCAN, enabling effective token mixing through self-attention in the spatial domain and facilitating the acquisition of multimodal representations. To effectively aggregate unimodal representations and cross-modal correlations, CAFE utilizes cross-modal alignment and disambiguation mechanisms. While it demonstrates good performance on the Weibo dataset, its effectiveness diminishes when applied to the Twitter dataset. BMR leverages multi-view learning to estimate the importance of different modalities for adaptive aggregated unimodal representation, resulting in superior performance. LogicDM considers logical relationships between predicates and selects predicates and cross-modal objects to derive and evaluate interpretable logical clauses, resulting in improved performance on the Twitter dataset.

Our proposed FSRU has delivered highly favorable results on both datasets, consistently ranking 1st or 2nd across all evaluation metrics. FSRU effectively explores and integrates multimodal features within the frequency domain. By leveraging the Fourier transform to bridge the spatial and frequency domains, FSRU achieves a lossless transformation of multimodal rumor features into a shared space. FSRU takes a cross-modal perspective to control spectral components while also capturing the intrinsic characteristics of rumors from an unimodal perspective. This conceptually straightforward yet computationally efficient approach significantly enhances the performance of rumor detection. In addition, FSRU employs multimodal feature aggregation based on distributional similarity and two types of con-

Datasets	Methods	Accuracy	Rumor			Non-rumor		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	att-RNN (Jin et al. 2017)	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN (Wang et al. 2018)	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	MVAE (Khattar et al. 2019)	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake (Singhal et al. 2019)	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	HMCAN (Qian et al. 2021)	0.885	0.920	0.845	0.881	0.856	**0.926**	0.890
	CAFE (Chen et al. 2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	BMR (Ying et al. 2023)	0.884	0.875	0.886	0.880	0.874	0.881	0.877
	LogicDM (Liu, Wang, and Li 2023)	0.852	0.862	0.845	0.853	0.843	0.859	0.851
	FSRU	**0.901***	**0.922***	0.892	0.906	**0.879***	0.913	**0.895***
Twitter	att-RNN (Jin et al. 2017)	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN (Wang et al. 2018)	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE (Khattar et al. 2019)	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake (Singhal et al. 2019)	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	HMCAN (Qian et al. 2021)	0.897	0.971	0.801	0.878	0.853	0.979	0.912
	CAFE (Chen et al. 2022)	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	BMR (Ying et al. 2023)	0.872	0.842	0.751	0.794	0.885	0.931	0.907
	LogicDM (Liu, Wang, and Li 2023)	0.911	0.909	0.816	0.859	**0.913**	0.958	0.935
	FSRU	**0.952***	**0.983***	**0.938***	**0.960***	0.901	**0.984***	**0.940***

Table 1: Performance comparison on the Weibo and Twitter datasets. The best performance is highlighted in bold, while underlining highlights the follow-up, and $*$ indicates the statistically significant improvement (i.e., two-sided t-test with $p < 0.05$).

trastive learning to learn the complementary relationships between cross-modal features. This allows FSRU to adaptively aggregate multimodal features for detection. However, it is important to note that the impact on the Weibo dataset appears to be slightly less pronounced compared to the Twitter dataset, possibly due to inherent differences between the two datasets. Firstly, the Weibo dataset is relatively smaller in size when compared to the Twitter dataset. Secondly, the Weibo dataset comprises a subset of images that exhibit lower quality or contain less informational content.

Ablation Study

To assess the effectiveness of different modules within FSRU, we conduct a comparative analysis with sub-models denoted as "-w/o USC", "-w/o CSC", "-w/o DSF", and "-w/o CL". These variants represent FSRU without considering unimodal spectrum compression, cross-modal spectrum co-selection, distribution similarity-based fusion, and dual contrastive learning, respectively. The results are shown in Table 2 and Figure 4.

Quantitative Analysis As shown in Table 2, it is evident that removing either the unimodal spectrum compression or the cross-modal spectrum co-selection adversely affects the

model's performance on both datasets. Without employing unimodal spectrum compression, the model loses the ability to explore distinctive patterns in modal frequency responses. Similarly, the absence of cross-modal spectrum component interactions hinders the model's capacity to learn dependencies between multimodal features. Moreover, excluding the distribution similarity-based fusion and the dual contrastive learning module from the model leads to a slight decline in performance. These findings highlight the significance of fusing multimodal features by measuring multimodal distribution similarity and leveraging dual contrastive learning.

Qualitative Analysis To further analyze the effect of the frequency spectrum representation and fusion module, we qualitatively visualize the features on the Weibo and Twitter test set with t-SNE as depicted in Figure 4 and Appendix E Figure **??**. The FSRU variants "-w/o USC" and "-w/o CSC" demonstrate the ability to discriminate multimodal rumor features, but there is a clear overlap between features across different labels. In contrast, the features learned by FSRU exhibit clear boundaries between labels, effectively reducing the overlapping between features.

Impact of the Number of Filters k

We conducted experiments by varying the value of k in USC from 1 to 8, as presented in Table 3. The results exhibit a pattern of initially increasing performance followed by a subsequent decline on both datasets. Specifically, there is a significant performance improvement from $k = 1$ to $k = 2$, while a slight decrease is observed from $k = 2$ to $k = 8$. By setting $k = 2$, the model has the ability to acquire diverse and distinct feature patterns from various dimensions of the frequency response while still maintaining an appropriate computational cost. Therefore, we determine that $k = 2$ is the optimal choice for FSRU on both datasets.

	Weibo		Twitter	
	Accuracy	F1	Accuracy	F1
FSRU	**0.901**	**0.902**	**0.952**	**0.950**
-w/o USC	0.866	0.865	0.910	0.908
-w/o CSC	0.883	0.882	0.924	0.922
-w/o DSF	0.876	0.875	0.947	0.943
-w/o CL	0.889	0.889	0.937	0.936

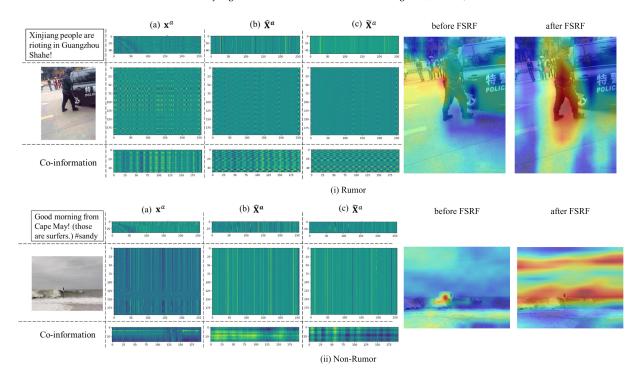Table 2: Comparison of different FSRU variants.

Figure 3: Interpretative visualization of rumor and non-rumor cases. Refer to Appendix D for more illustrative cases.

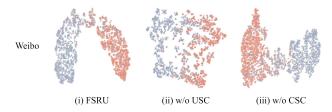Figure 4: T-SNE visualization of learned representations.

Filter	Weibo		Twitter	
	Accuracy	F1	Accuracy	F1
1	0.839	0.838	0.938	0.936
2	**0.901**	**0.902**	**0.952**	**0.950**
4	0.896	0.895	0.944	0.942
8	0.894	0.893	0.931	0.928

Table 3: Effect of the number of filters in USC.

Case Study

To provide an intuitive demonstration of the learning process of the Frequency Spectrum Representation and Fusion (FSRF) in FSRU, we visualize $\mathbf{x}^a$, $\hat{\mathbf{X}}^a$, and $\tilde{\mathbf{X}}^a$ ($a \in t, v$), along with the corresponding co-information for the two modalities, as shown in Figure 3. In the case of rumor, as FSRF is learned, the features gradually acquire a distinct pattern, allowing for better differentiation. This results in a clearer identification of concentrated spectral energy. On the other hand, in the case of non-rumors, the model seeks to capture truthfulness clues expressed through multimodal

features to the best of its ability. FSRF leverages co-selection across modalities to emphasize and suppress specific spectral features across modalities, thereby potentially revealing cues that indicate the veracity of rumors.

We have visualized the multimodal features of the two mentioned cases before and after the learning process of FSRF. In the first image, the model after FSRF learning concentrates on the person in the image, who does not match the person or event mentioned in the text. However, this person does not correspond to the individual or event mentioned in the accompanying text. This image therefore is classified as a rumor. In the second image, the model concentrates on the waves, the cloudy sky, and the surfer in the distance. This alignment between the visual elements and the textual description suggests consistency and coherence. Hence this image is classified as a non-rumor.

Conclusion

We first attempt to introduce a frequency spectrum representation and fusion network (FSRU) for multimodal rumor detection. FSRU is unique with a frequency spectrum representation and fusion to effectively capture both the frequency of feature changes and their intensity in the frequency domain, which is essential for FSRU to learn multimodal features properly. Substantial experiments demonstrate that our proposed approach achieves advanced performance. Our future studies include exploring deep insights and mechanisms in frequency-based multimodal fusion. The proposed model has the potential for more multimodal tasks and scenarios, we will further investigate the effectiveness and interpretability of the spectrum in multimodal fusion.

Acknowledgments

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Boididou, C.; Papadopoulos, S.; Kompatsiaris, Y.; Schifferes, S.; and Newman, N. 2014. Challenges of computational verification in social multimedia. In *WWW (Companion Volume)*, 743–748. ACM.

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; and Zhang, Q. 2020. Spectral Temporal Graph Neural Network for Multivariate Time-series Forecasting. In *NeurIPS*.

Chen, J.; Wu, Z.; Yang, Z.; Xie, H.; Wang, F. L.; and Liu, W. 2021. Multimodal Fusion Network with Latent Topic Memory for Rumor Detection. In *ICME*, 1–6. IEEE.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Lu, T.; and Shang, L. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *WWW*, 2897–2905. ACM.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast Fourier Convolution. In *NeurIPS*.

Cui, L.; Wang, S.; and Lee, D. 2019. SAME: sentiment-aware multi-modal embedding for detecting fake news. In *ASONAM*, 41–48. ACM.

Ehrlich, M.; and Davis, L. 2019. Deep Residual Learning in the JPEG Transform Domain. In *ICCV*, 3483–3492. IEEE.

Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers. *CoRR*, abs/2111.13587.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 9726–9735. Computer Vision Foundation / IEEE.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *ACM Multimedia*, 795–816. ACM.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *WWW*, 2915–2921. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization.

Koç, E.; and Koç, A. 2022. Fractional Fourier Transform in Time Series Prediction. *IEEE Signal Process. Lett.*, 29: 2542–2546.

Lange, H.; Brunton, S. L.; and Kutz, J. N. 2021. From Fourier to Koopman: Spectral Methods for Long-term Time Series Prediction. *J. Mach. Learn. Res.*, 22: 41:1–41:38.

LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10): 1995.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontañón, S. 2022. FNet: Mixing Tokens with Fourier Transforms. In *NAACL-HLT*, 4296–4313. Association for Computational Linguistics.

Li, S.; Xue, K.; Zhu, B.; Ding, C.; Gao, X.; Wei, D. S. L.; and Wan, T. 2020. FALCON: A Fourier Transform Based Approach for Fast and Secure Convolutional Neural Network Predictions. In *CVPR*, 8702–8711. Computer Vision Foundation / IEEE.

Lin, Z.; Liang, B.; Long, Y.; Dang, Y.; Yang, M.; Zhang, M.; and Xu, R. 2022. Modeling Intra- and Inter-Modal Relations: Hierarchical Graph Contrastive Learning for Multimodal Sentiment Analysis. In *COLING*, 7124–7135. International Committee on Computational Linguistics.

Liu, H.; Wang, W.; and Li, H. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. In *ACL (Findings)*, 9781–9796. Association for Computational Linguistics.

Mateos, G.; Segarra, S.; Marques, A. G.; and Ribeiro, A. 2019. Connecting the Dots: Identifying Network Structure via Graph Signal Processing. *IEEE Signal Process. Mag.*, 36(3): 16–43.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR (Workshop Poster)*.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *SIGIR*, 153–162. ACM.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global Filter Networks for Image Classification. In *NeurIPS*, 980–993.

Singhal, S.; Pandey, T.; Mrig, S.; Shah, R. R.; and Kumaraguru, P. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In *WWW (Companion Volume)*, 726–734. ACM.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *BigMM*, 39–47. IEEE.

Sun, M.; Zhang, X.; Ma, J.; and Liu, Y. 2021. Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection. In *EMNLP (Findings)*, 1412–1423. Association for Computational Linguistics.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *KDD*, 849–857. ACM.

Willmore, A. 2016. This analysis shows how viral fake election news stories outperformed real news on facebook.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.; and Ren, F. 2020. Learning in the Frequency Domain. In *CVPR*, 1737–1746. Computer Vision Foundation / IEEE.

Yang, L.; and Hong, S. 2022. Unsupervised Time-Series Representation Learning with Iterative Bilinear Temporal-Spectral Fusion. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 25038–25054. PMLR.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *CVPR*, 4084–4094. Computer Vision Foundation / IEEE.

Yi, K.; Zhang, Q.; Fan, W.; He, H.; Hu, L.; Wang, P.; An, N.; Cao, L.; and Niu, Z. 2023a. FourierGNN: Rethinking Multivariate Time Series Forecasting from a Pure Graph Perspective. *CoRR*, abs/2311.06190.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; Lian, D.; An, N.; Cao, L.; and Niu, Z. 2023b. Frequency-domain MLPs are More Effective Learners in Time Series Forecasting. *CoRR*, abs/2311.06184.

Yi, K.; Zhang, Q.; Wang, S.; He, H.; Long, G.; and Niu, Z. 2023c. Neural Time Series Analysis with Fourier Transform: A Survey. *CoRR*, abs/2302.02173.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping Multi-view Representations for Fake News Detection. In *Proceedings of the AAAI conference on Artificial Intelligence*.

Zhang, Q.; Yang, Y.; Shi, C.; Lao, A.; Hu, L.; Wang, S.; and Naseem, U. 2023. Rumor Detection With Hierarchical Representation on Bipartite Ad Hoc Event Trees. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zhang, T.; Wang, D.; Chen, H.; Zeng, Z.; Guo, W.; Miao, C.; and Cui, L. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In *IJCNN*, 1–8. IEEE.

Zheng, J.; Zhang, X.; Guo, S.; Wang, Q.; Zang, W.; and Zhang, Y. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. In *IJCAI*, 2413–2419. ijcai.org.