Fine-Grained Differences-Similarities Enhancement Network for Multimodal Fake News Detection

Xiaoyu Wu[1], Shi Li*[2], Zhongyuan Lai[3], Haifeng Song[4], Chunfang Hu[5]
College of Computer and Control Engineering, Northeast Forestry University, Harbin, China[1,2]
DeepVerse Technology (Shanghai) Ltd, ShangHai, China[3]
School of Electronics and Information Engineering, Taizhou University, Taizhou, China[4]
DeepVerse Technology (Shanghai) Ltd, ShangHai, China[5]

***Abstract*—The use of social media has proliferated dramatically in recent years due to its increasing reach and ease of use. Along with this enlarged influence of social media platforms and the relative anonymity afforded to content contributors, an increasingly significant proportion of social media is composed of untruthful or "fake" news. Hence for various reasons of personal and national security, it is essential to be able to identify and eliminate fake news sources. The automated detection of fake news is complicated by the fact that most news posts on social media takes very diverse forms, including text, images, and videos. Most existing multimodal fake news detection models are structurally complex and not interpretable; the main reason for this is the difficulty of identifying essential features which characterize fake social media posts, leading to different models focusing on multiple different aspects of the news detection task. In this paper, we show that contrasting the *different and similar* (DS) features of social media posts serves as an important identifying marker for their authenticity, with the consequence that we only need to direct our attention to this aspect when designing a multimodal fake news detector. To address this challenge, we propose the Fine-Grained Differences-Similarities Enhancement Network (FG-DSEN), which improves detection with a simple and interpretable structure to enhance the DS aspect between images and text. Our proposed method was evaluated on two different language social media datasets, Weibo in Chinese and Twitter in English. It achieved accuracies 3% and 3.8% higher than other state-of-the-art methods, respectively.**

***Keywords*—Fake news detection; social media; pre-training model; multimodal; transformer**

I. Introduction

Despite the challenges posed by the pandemic and current economic climate, virtual social media platforms have flourished in recent years, resulting in a significant increase in information volume. A large number of consumers tend to acquire news exclusively through social media platforms instead of conventional sources. However, most platforms lack timely and effective supervision, making it easy for people to publish unverified news [1]. Furthermore, in contrast to the linear dissemination mode of traditional media, social media's dissemination characteristics facilitate the rapid spread of information, directly resulting in explosive growth of disseminated fake news and subsequently broader negative impacts with greater potential for social harm. Early research on fake news often used a vague definition that could include hoaxes, satires, or clickbaits as fake news. This study describes fake news as "news articles that are intentionally and verifiably false and could mislead readers" [2].

Apart from its rates of spread, social media news items are often multimodal and include text, images, and videos. Posts containing multimodal information are more appealing to readers, and studies have shown that news with images is, on average, 11 times more probable to be shared than text-only news [3]. However, the multimodal of most social media news items poses a significant obstacle to fake news detection [4]. The determination of news veracity based solely on text or images brings a challenge. Therefore, the development of models capable of integrating multimodal information is crucial for effective fake news detection.

Fake news has been a concern for researchers for some years. Due to the increasing realization of fake news' effects on society, work on fake news detection has dramatically intensified in recent years [5], [6]. With multimodal features proving effective in enhancing detection rates, so has research on fake news detection increasingly focussed on multimodal integration. Most multimodal methods involve feature interactions that are very complex, with multiple overlapping fusion modules, with the consequence that the final models are often very involved and are not amenable to extensions or improvements. This also leads to the difficulty in explaining these models, and it tends to neglect a very important feature of fake news: the differences and similarities between the text and image components of fake news [7]. This, in turn, contributes to the suboptimal performance of fake news detection. An emphasis on considering these differences and similarities can highlight distinctions between fake and real posts, thereby improving the discriminative power of the final classifier.

Based on these considerations and in contrast to most prevailing fake news detection networks, our work uses a simple structure to extract similar and dissimilar information among different modalities to enhance features and improve detection performance. Our structure is similar to the co-attention transformer [8], [9], but the performance is better. It can establish connections between similar and dissimilar points among features of different modalities while ensuring the purity of the original information to achieve the best detection performance. Earlier studies either overlooked this crucial connection or employed overly intricate interaction modules that compromised the integrity of the original information. The detection accuracy is at least 3% higher than the best method on two datasets.In concrete terms, we introduce the *Fine-Grained Differences-Similarities Enhancement Network (FG-DSEN)* to detect fake news. The architecture comprises four main components: two fine-grained feature extractors, a differences-

similarities enhancement network, and a fake news classifier. Our work makes the following contributions:

- We propose an effective detection method for fake news that emphasizes the differences and similarities (DS) between true and fake news items in terms of their textual and visual modalities. We find that these predominant features serve as accurate evaluation metrics for truthfulness. Our method avoids overcomplicating the task by requiring only text and image information.
- To effectively detect fake news via DS detection, we design a deep learning model called FG-DSEN. The proposed method extracts fine-grained features using an optimized pre-trained model. The differences-similarities information of these features are enhanced and used for classification, allowing the model to achieve higher accuracies than the baselines.
- We evaluate on two standard datasets and is able to achieve state-of-the-art performance on both. For purposes of comparison we have included a significant number of SOTA baselines from very recent works. We also perform extensive ablation studies on our model and conclude that its efficacy is indeed due to emphasis on the difference-similarity contrast of extracted news features.

The paper is structured into five sections. Section I provides an introduction to the study. Section II summarizes early research in the field. Section III provides a detailed description of the proposed methodology. Section IV covers the experimental aspect, including datasets introduction, parameter settings, comparative and ablation experiments, as well as corresponding result analysis. Finally, Section V concludes the work and discusses future directions.

II. Related Work

This study investigates approaches to detecting fake news using machine learning and deep learning methods, and additionally categorizing the latter into single-and multimodal methods.

A. Machine Learning-based Fake News Detection

Early works on fake news detection typically employed hand-crafted features combined with machine learning models. Among the widely studied hand-crafted features are text-specific content features such as punctuation [10], [11], [12], textual sentiment polarity [3], [13], and personal pronouns [3], [10], [11]. Propagating features, such as the root degree in the propagation tree and the average degree of leaf nodes [3], [10], are significant indicators. User features like user's following and followed [3], [10], [12], [14] and account profile completeness [15]. These hand-crafted features can then be utilized to train machine learning models, including decision trees and SVMs. Castillo et al. [10] crawled hot news on Twitter for approximately two months and constructed a decision tree for news credibility determination. Pérez-Rosas et al. [11] manually curated features related to language, including punctuation, psycholinguistic features, and syntactic rules. These methods necessitate researchers to possess extensive knowledge of linguistics and to know which features effectively distinguish true from false news, which is a difficult task.

B. Deep Learning-based Fake News Detection

1) Single-modal approaches to fake news detection: With the advent of deep learning models, researchers have discovered that they outperform traditional machine learning models and no longer necessitate intricate hand-crafted features. Kaliyar et al. [16] introduced a deep convolutional neural network (FNDNet) for fake news detection, which takes as input a word embedding vector produced by GloVe. Sahoo et al. [17] present a technique for detecting the authenticity of news content by combining user profiles. The exBAKE model presented by Jwa et al. [18] uses weighted cross-entropy to classify the data, which mitigates the issue of data imbalance in BERT [19].

2) Multimodal approaches to fake news detection: News images contain abundant information, so an increasing number of researchers are focusing on fake news detection methods that fuse multimodal features. Singhal et al. [20] proposed Spotfake, a multimodal approach for detecting fake news, which employs VGG-19 to extract image features and BERT to extract textual features. Wang et al. [21] proposed an Event Adversarial Neural Network (EANN) that utilizes event discrimination as an auxiliary task. The att-RNN framework proposed by Jin et al. [22] leverages an attention mechanism to enhance modal information, integrates social context features, and feeds the fused multimodal information into the classifier for classification. Wu et al. [23] proposed a Multimodal Co-attentive Networks (MCAN), and Qian et al. [24] proposed a Hierarchical Multi-modal Contextual Attention Network (HMCAN). Both approaches utilize transformer-based attention modules to combine features from various modalities or layers, thereby aiding detection. Jing et al. [25] presented a Multimodal Progressive Fusion Networks (MPFN) to retain shallow information by sampling and fusing features at different levels.Although previous works have achieved superior results in fake news detection tasks, they neglected the extraction of fine-grained features and mapping alignment and failed to exploit similarities and differences of features when fusing in depth.

III. Methodology

A. Model Overview

The present study establishes the task of detecting fake news as follows: given a multimodal news post $S = (T, V)$ containing text T and images V. Training a model then implies a mapping $f : S \rightarrow Y$, where our predefined classes $Y \in \{0, 1\}$ with $Y = 0$ implying that a news post S is fake and $Y = 1$ implying that it is true. More details about the model will be provided in the later subsections.

The proposed FG-DSEN is shown in Fig. 1. The model consists of

- Two fine-grained feature extractors: We utilize the VGG-19 [26] and 1D convolution to extract fine-grained features from images; for text we employ BERT and a BiLSTM;

- A differences-similarities enhancement network: Two self-attention modules are applied in parallel to extract feature similarities and differences. The attention module is able to capture correlation between these;
- A fake news classifier: We connect a simple fully-connected feedforward network to perform classification of the fused features.

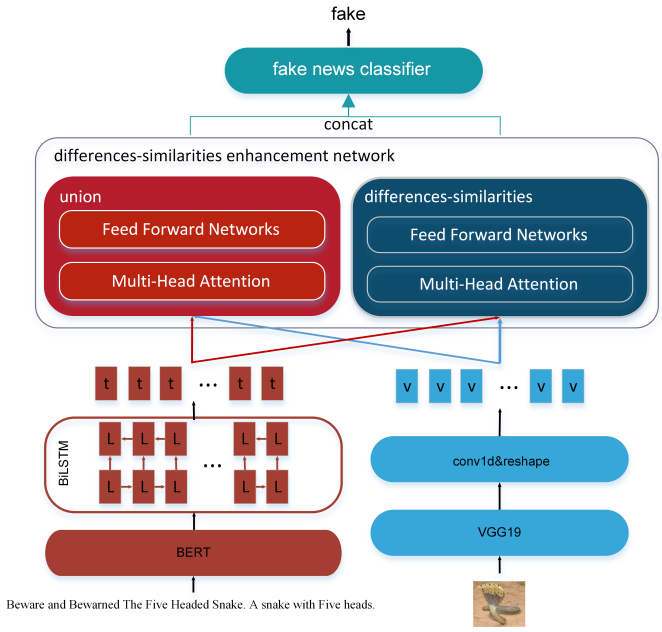


Fig. 1. The proposed model architecture.

B. Text Feature Extractor

Fake news detection conventionally employs static word embedding models.In our case, in order to context information explicitly into account, we use a pre-trained BERT model which has excellent performance in dynamic word embedding tasks. Texts T of social media news posts have the general structure $T_i = [x_1, x_2, \ldots, x_n]$, where $x_j \in T_i$, $j \in \{1, 2, \ldots, n\}$ is the jth word in the ith text corpus. We then obtain the dynamically extracted word vector, representing the fine-grained word-level features, via a BERT model. These are input into a BiLSTM to generate the text fine-grained feature F^T containing global text information, as shown in Eq. (1).

$$F^T = \text{BiLSTM}\left(\text{BERT}\left(t_i; \Theta_{\text{BERT}}\right); \Theta_{\text{BiLSTM}}\right), \quad (1)$$

where Θ_{BERT} and Θ_{BiLSTM} are the BERT and BiLSTM model parameters, respectively.

C. Image Feature Extractor

We utilize the VGG-19 as an image feature extractor by eliminating the average pooling and classification layers. We also remove the final 2D Max-pooling layer in the feature extraction part to reduce the loss of image information and facilitate alignment with text features. Image i_j of news posts is input to the modified VGG-19 to obtain the fine-grained pixel-level features.

$$\mathbf{f}_{\text{VGG-19}} = \text{VGG-19}\left(V_j; \Theta_{\text{VGG-19}}\right) \quad (2)$$

Subsequently, these are aligned with text fine-grained features through reshaping and 1D convolution to obtain fine-grained image features, as shown in Eq. (3).

$$F^V = \texttt{Conv1d}\left(\texttt{Reshape}(\mathbf{f}_{\text{VGG-19}})\right). \quad (3)$$

D. Differences-Similarities (DS) Enhancement Network

A differences-similarities (DS) enhancement network was designed based on ideas on feature processing involving fusion of fine-grained feature similarities and differences before classification [25]. Firstly, F^T and F^V are concatenated, as shown in Eq. (4). The direct concatenation of fine-grained features at the word-level and pixel-level is equivalent to modal alignment, hence establishing a link between word vector and feature map. Subsequently, a fully connected layer with the `LeakyReLU` activation function is applied to map the concatenated features into the same semantic space, as shown in Eq. (5).

$$F^{VT} = \texttt{Concat}(F^T, F^V) \quad (4)$$

$$P^{VT} = \texttt{max}(0, W_F F^{VT}) + 0.01 * \texttt{min}(0, W_F F^{VT}) \quad (5)$$

where F^{VT} represents the stitching feature, P^{VT} represents the stitching projection feature, and W_F denotes the fully connected layer's weight. F^T and F^V are connected after subtracting and finding the Hadamard product, respectively, as shown in Eq. (6). Subtraction yields the different points between image and text features, while the Hadamard product amplifies the similar points of image and text features. Then a fully connected layer with the `LeakyReLU` activation function maps text and image differences-similarities features into the same semantic space, as shown in Eq. (7).

$$FC^{VT} = \texttt{Concat}((F^T - F^V), (F^T \odot F^V)) \quad (6)$$

$$PC^{VT} = \texttt{max}(0, W_{FC} FC^{VT}) + 0.01 * \texttt{min}(0, W_{FC} FC^{VT}) \quad (7)$$

where FC^{VT} represents the differences-similarities feature, PC^{VT} represents the differences-similarities projection feature, and W_F denotes the fully connected layer's weight.

After deep-fusing P^{VT} through the self-attention module, we obtain the union attention feature S^U. Similarly, PC^{VT} is passed through another self-attention module to obtain the DS attention feature S^C. The DS attention feature is used to help distinguish fake news by interacting with the similar and dissimilar parts of the two modalities. The union attention feature then ensures that the original multimodal information is not lost after enhancement by the differences-similarities enhancement network to help the model comprehend the whole news. An example of obtaining a union attention feature is shown in Eq. (8)-(11).

$$S_1^U = \texttt{Softmax}\left(\frac{Q_U K_U^T}{\sqrt{d_k}}\right) V_U \quad (8)$$

$$S_2^U = \texttt{Layer\_norm}\left(P^{VT} + S_1^U\right) \quad (9)$$

$$FFN\left(S_2^U\right) = \texttt{max}\left(0, W_1 S_2^U\right) W_2 \quad (10)$$

$$S^U = \texttt{Layer\_norm}(S_2^U + FFN\left(S_2^U\right)) \quad (11)$$

where $Q_U = P^{VT} W_Q$, $K_U = P^{VT} W_K$, $V_U = P^{VT} W_V$, This indicates that after inputting P^{VT}, the Q_U, K_U, V_U required to calculate the self-attention are obtained by multiplying with

their respective weight matrices W_Q,W_K,W_V. The denominator in Eq. (8) is a scale factor controlling the magnitude of the attention fraction, where d_k represents the dimension of Q. W_1 and W_2 are the two weight matrices in the position-wise feed-forward networks. `Layer_norm` refers to layer normalization. Finally, S^U and S^C are reconnected after mean-pooling to obtain the enhanced feature S^{UC}, as shown in Eq. (12).

$$S^{UC} = \texttt{Concat}(\texttt{MeanPooling}(S^U), \texttt{MeanPooling}(S^C)) \tag{12}$$

E. Fake News Classifier

The fake news classifier employed in this study is a MLP consisting of two fully connected layers. The association between characteristics and classifications is accomplished via the softmax activation function applied to the output layer of the MLP upon feeding S^{UC}. This is demonstrated in Eq. (13) and (14).

$$MLP_1 = \texttt{max}(0, W_{s1}S^{UC}) + 0.01 * \texttt{min}(0, W_{s1}S^{UC}) \tag{13}$$

$$P = \texttt{Softmax}\,(W_P MLP_1 + b_P) \tag{14}$$

In the fully connected layer above, W_{s1} and W_P represent the weight matrix and b_P represents the bias term. This study uses cross-entropy as the loss function.

IV. Experiments

This section presents the results of numerical experiments performed on our proposed model, evaluated on two well-known social media datasets (Weibo and Twitter). We introduce these datasets and discuss some state-of-the-art baseline models. Then FG-DSEN is compared with these methods and we show that it achieves SOTA results compared to the baselines. Finally we present results from ablation experiments and analyze and interpret the findings derived from our experiments.

A. Datasets

- Twitter: The dataset [27] contains the text of news posts, additional images, and corresponding IDs, and includes a development and a test set. The former comprises about 5,000 real and 6,000 fake news posts. The latter contains about 2,000 news posts. We retained only those news samples that contained both text and images and used Google Translate to standardize the language of the tweets to English.
- Weibo: Jin et al. [22] collected and published this dataset. Fake news posts were sourced from all news articles published by the official Weibo platform disinformation system between May 2012 and January 2016, which had been verified as fake. The system enlists reputable users to review tweets reported by regular users to determine their veracity. The real news posts were sourced from posts verified by the official Xinhua News Agency. During dataset processing, we remove news samples containing only text or images in the dataset and eliminate duplicate or low-quality images. Table I presents detailed statistics for both datasets.

TABLE I. Statistics of Two Datasets

Dataset	Label	Number	Total
Weibo	Fake	4749	9528
	Real	4779	
Twitter	Fake	7021	12995
	Real	5974	

TABLE II. Different Hyperparameters on Two Datasets

Hyperparameter	Value	
	Weibo	Twitter
Sentence length	192	87
BERT vision	BERT_base_chinese	BERT_base_uncased
Minibatch size	70	128
Epoch	150	100
Learning rate	0.00005	0.0001

B. Data Preprocessing and Experimental Settings

Text sequences were converted to dynamic word vectors using BERT-base with a dimension of 768. For image data, we resized images to $224 \times 224 \times 3$ and fed them to the modified VGG-19 model for fine-grained feature extraction, yielding a dimensionality of 100352. We froze the parameters of both BERT and VGG-19 models to prevent overfitting. The BiLSTM had a dropout rate of 0.4 and a dimension of 256; the convolution kernel size of the 1D convolution is set to 1, with a stride of 1 and an output channel count equal to the length of the text sequence. The two transformer encoders of the self-attention module are identical, with dimensions of 256, 8 attention heads, and a dropout rate of 0.4. We optimize parameters utilizing the Adam optimizer. The hyperparameters that were different on the two datasets during training are presented in Table II.

C. Evaluation

We compare FG-DSEN with other single- and multimodal methods to evaluate its performance on fake news detection tasks.

1) Single-modal based approaches:

- Text: We omit the image embedding component from the FG-DSEN. After extracting text fine-grained features using BERT and BiLSTM and performing differences-similarities feature extraction, we input to a self-attention module and perform fake news classification via a MLP layer;
- Images: Similar to the processing of text-only news classification, we exclude the text embedding layer from the FG-DSEN and use VGG-19 and a Conv1d to extract fine-grained image features before directly feeding them into a self-attention module and MLP for classification.

2) Multimodal-based approaches:

- Att_RNN [22]: Att_RNN is a recurrent neural network incorporating an attention mechanism for modal fusion in rumor detection. To ensure a fair comparison

with our approach, we utilize only text and image features in our experiments and remove the components that deal with social context information;

- EANN [21]: Event Adversarial Neural Networks(EANN) extract event-invariant features by adding an event discriminator as a secondary task to help better detect fake news. For our experiments, we utilize EANN with the event discriminator component removed;
- MVAE [28]: The Multimodal Variational Autoencoder (MVAE) consists of a bimodal variational autoencoder and a binary classifier. It employs pre-trained VGG-19 and BiLSTM to mine features from images and text;
- SpotFake [20]: SpotFake is a multimodal framework developed for the detection of fake news. The framework employs VGG-19 to extract image features and a pre-trained language model, BERT, to extract text features;
- SpotFake+ [29]: SpotFake+ is built on top of SpotFake, which utilizes a pre-trained XLNet model instead of BERT to extract text features and employs richer fully-connected layers to assist in modal fusion;
- SAFE [30]: SAFE is a multimodal fake news detection approach based on perceptual similarity. The framework introduces auxiliary objective functions to measure text and image similarity, aiding in detecting fake news by incorporating measures beyond simply splicing multimodal features together;
- HMCAN [24]: Hierarchical Multimodal Contextual Attention Network (HMCAN) captures hierarchical semantic information through a hierarchical coding network. Multimodal contextual attention networks are used to fuse inter-modality and intra-modality relationships;
- MPFN [25]: the Multimodal Progressive Fusion Network (MPFN) uses Swin Transformer to extract multilevel visual features from images, VGG-19 to extract additional frequency domain features from images, and BERT to extract text features.

D. Results and Analysis

We conducted broad experiments on two public datasets to evaluate our model's effectiveness and generalization ability. Table III shows that the overall performance of the FG-DSEN surpasses that of the baseline approach. Based on these results, the following conclusions can be drawn:

- For both single-modal methods, neither perform as well as the original FG-DSEN. However, the text single-modal method's accuracy in the Weibo dataset surpasses all other multimodal methods except our proposed method. This demonstrates that text fine-grained features with the self-attention module are highly effective for news classification. Its accuracy in the Twitter dataset is second only to HMCAN, probably because the text in the Twitter dataset is short, and some of the text needs to be translated with high quality, which impacts performance. The image single-modal approach has the lowest accuracy on the Weibo dataset. At the same time, it outperforms all multimodal approaches except our proposed method in the Twitter dataset, which proves that fine-grained image features with the self-attention module can sometimes be very effective;
- Both att-RNN and EANN methods exhibit diminished performance after excluding additional social background information and auxiliary tasks. This indicates that incorporating auxiliary tasks or additional features can enhance fake news detection performance. Nevertheless, the overall effect falls short of that achieved by a model designed specifically for various multimodal feature fusion;
- The inferior performance of MVAE compared to SpotFake demonstrates that the improvement brought by auxiliary tasks is not as effective as using pre-trained models. The fact that SpotFake is less effective than SpotFake+ suggests that using a better pre-trained model can enhance fake news detection;
- MPFN and HMCAN extract image and text features hierarchically and design a complex fusion network for hierarchical feature fusion abd therefore better utilizing shallow-level features. However, multiple complex fusion networks increase the computational cost and do not focus on the similarities and differences between different modal features, resulting in suboptimal detection of fake news;
- The Precision, Recall, and F1-Score of real and fake news on the Weibo dataset are equal for the FG-DSEN; we investigate them by confusion matrix, as presented in Table IV. We can see that the cause of the equivalence is that false positives and negatives happen to be equal;
- Our proposed method's overall performance on the Weibo and Twitter datasets surpasses other baselines, and additionally, We have a simple structure with fewer parameters to train. Therefore, our method extracts fine-grained features to better and more efficiently capture the images and text information in the news. The differences-similarities attention feature can better extract each modality's similar and dissimilar information. The union attention feature can ensure the fusion of the original multimodal information.

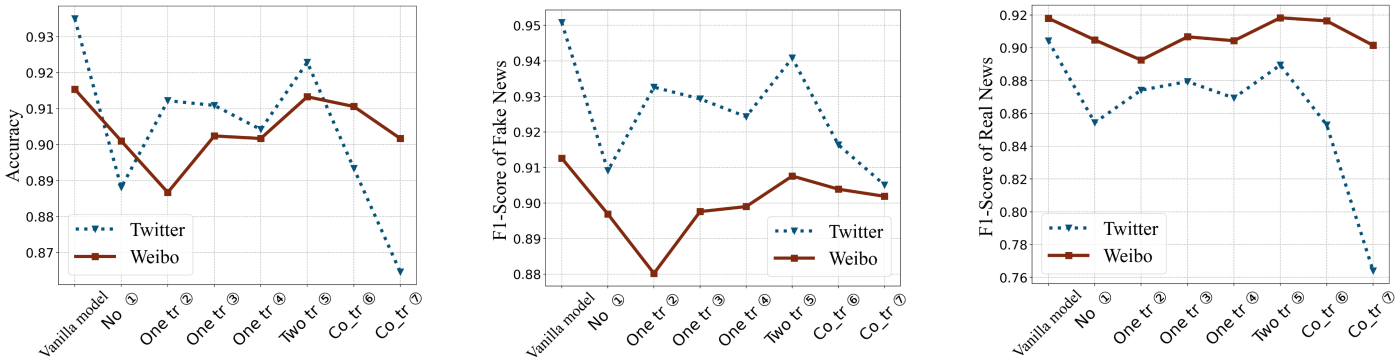
E. Model Ablation

This section presents ablation experiments conducted on FG-DSEN and compares them with a variant using the co-attention transformer to determine its effectiveness.

- No transformer: The stitching projection features and differences-similarities projection features are connected and input to the fake news classifier for experiment ①;
- One transformer: We utilize a single self-attention module within the DS enhancement network. Three distinct experiments are conducted: ② inputting only

TABLE III. THE RESULTS OF DIFFERENT METHODS ON WEIBO AND TWITTER DATASET. THE HIGHEST SCORE IS HIGHLIGHTED IN BOLD

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
WEIBO	Textual only	0.898	0.875	**0.920**	0.897	**0.921**	0.877	0.898
	Visual only	0.624	0.651	0.480	0.552	0.609	0.759	0.676
	att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MAVE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	**0.920**	0.845	0.881	0.856	**0.926**	0.890
	MPFN	0.838	0.857	0.894	0.889	0.873	0.863	0.876
	FG-DSEN	**0.915**	0.913	0.913	**0.913**	0.918	0.918	**0.918**
TWITER	Textual only	0.867	0.892	0.911	0.902	0.811	0.776	0.793
	Visual only	0.910	0.886	**0.999**	0.939	**0.997**	0.738	0.848
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	0.897	**0.971**	0.801	0.878	0.853	**0.979**	**0.912**
	MPFN	0.833	0.846	0.921	0.880	0.809	0.721	0.740
	FG-DSEN	**0.935**	0.965	0.937	**0.951**	0.879	0.931	0.904



(a) (b) (c)

Fig. 2. Depicts the accuracy and F1-score of FG-DSEN and its variants on two datasets. The overall impact of the differences-similarities enhancement network is demonstrated.

TABLE IV. CONFUSION MATRIX OF OPTIMAL RESULTS OF FG-DSEN ON THE WEIBO DATASET

Confusion Matrix		Y true	
		Positive	Negative
Y predicted	Positive	694	62
	Negative	62	647

stitching projection features; ③ inputting only DS projection features; and ④ inputting features connected by stitching projection features and DS projection features;

- Two transformers: Within the DS enhancement network, we use two self-attention modules with shared weights for the experiment ⑤, otherwise identical to the original network;
- Co-attention transformer: Substitute the two parallel self-attention module structures within the DS enhancement network with a single co-attention transformer. We conduct two experiments: ⑥ inputting the stitching projection features and DS projection features into the co-attention transformer, respectively, and ⑦ inputting the text fine-grained features and image fine-grained features directly into the co-attention transformer after projecting them to the same dimen-

TABLE V. RESULTS OF ABLATION EXPERIMENTS OF FG-DSEN

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
WEIBO	FG-DSEN	**0.9154**	0.9126	0.9126	**0.9126**	0.9180	0.9180	0.9180
	No ①	0.9010	0.9040	0.8900	0.8969	0.8983	0.9114	0.9048
	One tr ②	0.8867	0.9010	0.8604	0.8802	0.8744	0.9114	0.8925
	One tr ③	0.9024	0.9113	0.8843	0.8976	0.8945	0.9193	0.9067
	One tr ④	0.9017	0.8940	0.9041	0.8990	0.9091	0.8995	0.9043
	Two tr ⑤	0.9133	0.9369	0.8801	0.9076	0.8936	0.9444	**0.9183**
	Co_tr ⑥	0.9106	**0.9419**	0.8688	0.9039	0.8853	**0.9497**	0.9164
	Co_tr ⑦	0.9017	0.8722	**0.9337**	0.9019	**0.9334**	0.8717	0.9015
TWITTER	FG-DSEN	**0.9350**	0.9650	0.9370	**0.9508**	0.8791	0.9309	**0.9042**
	No ①	0.8881	**0.9976**	0.8305	0.9091	0.7481	**0.9959**	0.8544
	One tr ②	0.9122	0.9608	0.9060	0.9326	0.8288	0.9248	0.8742
	One tr ③	0.9109	0.9921	0.8740	0.9293	0.7938	0.9858	0.8794
	One tr ④	0.9042	0.9820	0.8730	0.9243	0.7894	0.9675	0.8694
	Two tr ⑤	0.9229	0.9692	0.9140	0.9408	0.8434	0.9411	0.8895
	Co_tr ⑥	0.8934	0.9657	0.8720	0.9164	0.7827	0.9370	0.8529
	Co_tr ⑦	0.8646	0.8537	**0.9630**	0.9051	**0.8984**	0.6646	0.7640

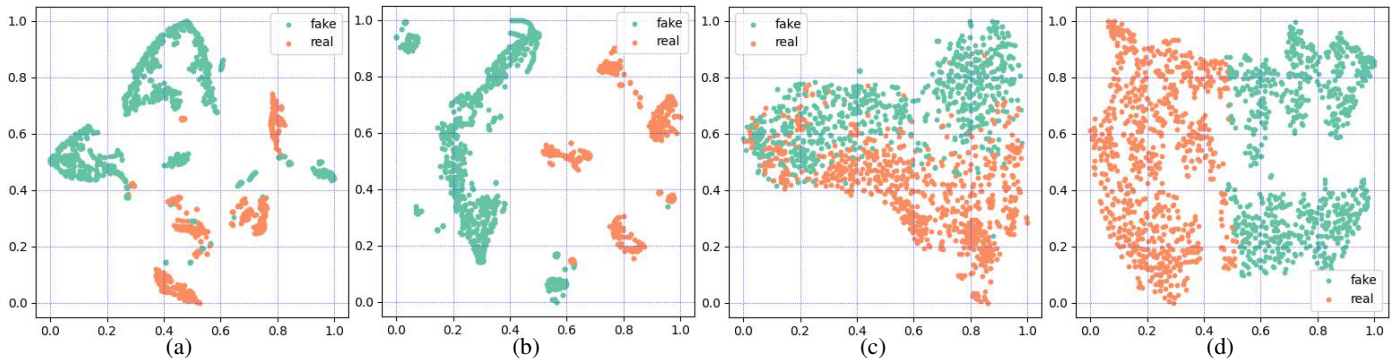


Fig. 3. (a) and (b) depict two features from the Twitter dataset, while (c) and (d) depict two features from the Weibo dataset. (a) and (c) represent features that have not undergone differences-similarities enhancement network processing, while figures (b) and (d) represent those that have.

sion as the stitching projection features.

Table V and Fig. 2 present the results of these ablation experiments.

The overall impact of the DS enhancement network is demonstrated by comparing the features before and after the network on the two datasets, see Fig. 3. We can find that after passing through the differences-similarities enhancement network, samples from each category are more closely clustered and exhibit distinct boundaries.

Based on Experiment ①, it is evident that the classification performance deteriorates in the absence of the DS enhancement network. Within the DS enhancement network, we compare experiments ②, ③, and ④ and find that only DS projection features work best on the Weibo dataset, and only stitching projection features work best on the Twitter dataset. This indicates that each of the two features is effective. However, directly connecting stitching projection features and DS projection features and inputting them into the self-attention module for enhancement proves to be ineffective. A possible explanation is that the two features are not in the same semantic space and represent original and DS information. Excessive fusion can have a detrimental effect and contaminate the extracted features. Experiment ⑤ indicates a slight decrease in model performance due to a reduction in the number of parameters. However, the shared weight model was the most effective in the ablation experiments, except for the original model, and outperformed all other baseline models in the previous subsection in performance. From experiments ⑥, ⑦, we can see that two self-attention modules in parallel work much better than a single co-attention transformer, despite having the same number of parameters. Additionally, directly inputting text fine-grained features and image fine-grained features of the same dimension does not effectively capture DS information in news content and is not conducive to detecting fake news. These experiments validate the rationality of our designed DS enhancement network.

V. Conclusion

This study proposes a fine-grained differences-similarities (DS) enhancement network for fake news detection. Initially, the model extracts fine-grained features from text and images using a modified pre-training model to minimize the loss of valid information in news posts. The DS enhancement network is then employed to enhance both the stitching features and DS features of the news, ensuring interaction between original modal information while highlighting DS information in news content to aid in fake news detection. The improvement achieved by using the DS enhancement network is more significant than that achieved by adding auxiliary tasks or extracting multi-level features. Experiments demonstrate that the FG-DSEN proposed in this paper outperforms state-of-the-art methods on two public and popular social media datasets.

This study has certain limitations, as it cannot be directly applied when one of the modalities is missing. In scenarios where news articles solely comprise of textual or visual content, models exhibit limited performance in handling such cases. Additionally, if manipulated images or artificially generated deceptive images are employed to align with fake news narratives, it is possible that the extraction of DS features could be compromised, leading to potential challenges in detecting fake news.

For future work, we aim to enhance the model's capabilities in the detection of fake news in scenarios where modalities are missing. This would enable the model to be more versatile in detecting both single-modal and multimodal fake news. Furthermore, we plan to augment the feature extraction process to address manipulated images or artificially generated deceptive images, incorporating them into the construction of DS features. This enhancement is expected to fortify the model's robustness and improve its overall performance.

References

[1] C. G. Song, N. W. Ning, Y. L. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, p. 14, 2021.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] Z. W. Jin, J. Cao, Y. D. Zhang, J. S. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *Ieee Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[4] C. Raj and P. Meel, "Convnet frameworks for multi-modal fake news detection," *Applied Intelligence*, pp. 1–17, 2021.

[5] T. Rasool, W. H. Butt, A. Shaukat, and M. U. Akram, "Multi-label fake news detection using multi-layered supervised learning," in *Proceedings of the 2019 11th international conference on computer and automation engineering*, Conference Proceedings, pp. 73–77.

[6] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, Conference Proceedings, pp. 1–8.

[7] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, and G. Xu, "Entity-oriented multi-modal alignment and fusion network for fake news detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 3455–3468, 2021.

[8] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in neural information processing systems*, vol. 32, Conference Proceedings, pp. 13–23.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, Conference Proceedings, pp. 5998–6008.

[10] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, Conference Proceedings, pp. 675–684.

[11] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2019, Conference Proceedings, pp. 3391–3401.

[12] V. Sahana, A. R. Pias, R. Shastri, and S. Mandloi, "Automatic detection of rumoured tweets and finding its origin," in *2015 International Conference on Computing and Network Communications (CoCoNet)*. IEEE, Conference Proceedings, pp. 607–612.

[13] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security & Privacy*, vol. 1, no. 1, p. e9, 2018.

[14] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, "A credibility analysis system for assessing information on twitter," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 661–674, 2018.

[15] V. Indu and S. M. Thampi, "A nature-inspired approach based on forest fire model for modeling rumor propagation in social networks," *Journal of network computer applications*, vol. 125, pp. 28–41, 2019.

[16] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[17] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, 2021.

[18] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. J. A. S. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," vol. 9, no. 19, p. 4062, 2019.

[19] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[20] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, Conference Proceedings, pp. 39–47.

[21] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "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*, Conference Proceedings, pp. 849–857.

[22] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, Conference Proceedings, pp. 795–816.

[23] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Conference Proceedings, pp. 2560–2569.

[24] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Conference Proceedings, pp. 153–162.

[25] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information Processing & Management*, vol. 60, no. 1, p. 103120, 2023.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015, pp. 1–14.

[27] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2015," in *MediaEval 2015 Workshop*, vol. 3, Conference Proceedings, p. 7.

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

[29] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, Conference Proceedings, pp. 13 915–13 916.

[30] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 2020, pp. 354–367.