applied sciences

MDPI

Article

FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection

Jingzi Wang, Hongyan Mao * and Hongwei Li *

Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China; jingziwang@163.com

* Correspondence: hymao@sei.ecnu.edu.cn (H.M.); 51194501008@stu.ecnu.edu.cn (H.L.)

Abstract: As one of the most popular social media platforms, microblogs are ideal places for news propagation. In microblogs, tweets with both text and images are more likely to attract attention than text-only tweets. This advantage is exploited by fake news producers to publish fake news, which has a devasting impact on individuals and society. Thus, multimodal fake news detection has attracted the attention of many researchers. For news with text and image, multimodal fake news detection utilizes both text and image information to determine the authenticity of news. Most of the existing methods for multimodal fake news detection obtain a joint representation by simply concatenating a vector representation of the text and a visual representation of the image, which ignores the dependencies between them. Although there are a small number of approaches that use the attention mechanism to fuse them, they are not fine-grained enough in feature fusion. The reason is that, for a given image, there are multiple visual features and certain correlations between these features. They do not use multiple feature vectors representing different visual features to fuse with textual features, and ignore the correlations, resulting in inadequate fusion of textual features and visual features. In this paper, we propose a novel fine-grained multimodal fusion network (FMFN) to fully fuse textual features and visual features for fake news detection. Scaled dot-product attention is utilized to fuse word embeddings of words in the text and multiple feature vectors representing different features of the image, which not only considers the correlations between different visual features but also better captures the dependencies between textual features and visual features. We conduct extensive experiments on a public Weibo dataset. Our approach achieves competitive results compared with other methods for fusing visual representation and text representation, which demonstrates that the joint representation learned by the FMFN (which fuses multiple visual features and multiple textual features) is better than the joint representation obtained by fusing a visual representation and a text representation in determining fake news.

Keywords: fake news detection; feature fusion; attention mechanism; social media

check for updates

Citation: Wang, J.; Mao, H.; Li, H. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Appl. Sci.* **2022**, *12*, 1093. https://doi.org/10.3390/app12031093

Academic Editors: Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 13 December 2021
Accepted: 17 January 2022
Published: 21 January 2022

1. Introduction

With the rapid development of social networks, social media platforms have become ideal places for news propagation [1]. Due to its convenience, people are increasingly seeking out and consuming news through social media. However, the convenience also facilitates the rapid spread and proliferation of fake news [2], which has a devasting impact on individuals and society [3].

As one of the most popular social media platforms, microblogs, such as Twitter and Weibo, allow people to share and forward tweets, where the tweets with both text and images are more likely to attract attention than the text-only tweets. This advantage is also exploited by fake news producers, who post tweets about fake news on microblogs by manipulating text and forging images. If these tweets are not verified, they may seriously jeopardize the credibility of microblogs [4]. Therefore, it is crucial to detect fake news on microblogs.

In recent years, methods for fake news detection have gradually evolved from unimodal to multimodal approaches. The question concerning how to learn a joint representation that contains multimodal information has attracted much research attention. Jin et al. [4] use local attention mechanism to refine the visual representation, but the refined visual representation cannot reflect the similarity between the visual representation and the joint representation of text and social context. Wang et al. [5] propose a model based on adversarial networks to learn an event-invariant feature. Khattar et al. [6] propose a model based on variational autoencoder (VAE) to learn a shared representation. However, these models view the concatenation of unimodal features as a joint representation, which cannot discover dependencies between modalities. Song et al. [7] leverage an attention mechanism to fuse a number of word embeddings and one image embedding to obtain fused features, and further extract key features from the fuse features as a joint representation. Although the joint representation captures the dependencies, the fusion is not fine-grained enough. This is due to the fact that they do not use multiple feature vectors representing different visual features to fuse with textual features, and ignore correlations between different visual features.

To overcome the limitations of the aforementioned methods, the **f**ine-grained **m**ultimodal **f**usion **n**etworks (FMFN) is proposed for fake news detection. Our approach includes the following three steps. First, we use deep convolutional neural networks (CNNs) to extract multiple visual features of a given image and RoBERTa [8] to obtain deep contextualized word embeddings of words, each of which can be considered as a textual feature. Then, the scaled dot-product attention [9] is employed to enhance the visual features as well as the textual features, and fuse them. Finally, the fused feature is fed into a binary classifier for the detection.

The contributions can be summarized as follows:

1. To effectively detect fake news with text and image, we propose a novel model for fine-grained fusion of textual features and visual features.
2. The proposed model utilizes attention mechanism to enhance the visual features as well as the textual features, and fuse the enhanced visual features and the enhanced textual features, which not only considers the correlations between different visual features but also captures the dependencies between textual features and visual features.
3. We conduct extensive experiments on the real-word dataset. The results demonstrate the effectiveness of the proposed model.

This paper is organized as follows. In the next section, we review related work on fake news detection and scaled dot-product attention. Section 3 provides details of the proposed model. Section 4 presents the experiments. Section 5 gives the ablation analysis. In Section 6, we conclude the paper with a summary and give an outlook on future work.

2. Related Work

Fake news is defined as the news that is deliberately fabricated and is verifiable false [10,11]. Existing work on fake news detection can be divided into two categories: unimodal and multimodal. Scaled-dot product attention has been applied to the fields of natural language processing (NLP) and computer vision (CV). In NLP and CV, the extraction of corresponding features, such as textual features and visual features, is a fundamental task, and it is also a key step in fake news detection. In this section, we review the related work on unimodal fake news detection, multimodal fake news detection, and the scaled dot-product attention.

2.1. Unimodal Fake News Detection

Only one modality of content is utilized for unimodal fake news detection, such as text content, visual content, and social context. The text content of news plays an important role in determining the authenticity of the news. Ma et al. [12] use RNN to learn text representations from text content. Yu et al. [13] propose a CNN-based method to extract local-global significant features of text content. The two methods concentrate on detecting fake news at the event level, and thus require event labels, which increases the cost of the detection. To learn a stronger indicative representation of rumors, a GAN-style model is proposed by Ma et al. [14]. Besides text content, image is also crucial, which has a great influence on news propagation [15,16]. Qi et al. [17] use RNN and CNN-RNN to extract visual features in the frequency domain and the pixel domain, respectively. The visual features in different domains are then fused using an attention mechanism. In addition to textual features and visual features, social context features are also widely used for fake news detection on social media. To capture propagation patterns of news, Wu et al. [18] develop an SVM classifier based on kernel methods, which combine some social context features. For early detection of fake news, Liu et al. [19] extract user characteristics from user profiles to judge the authenticity of the news.

2.2. Multimodal Fake News Detection

Multimodal fake news detection relies on multimodal information, rather than information from one modality of content. The process involves feature extraction and feature fusion. In feature extraction, textual feature extractors can be implemented using Bi-LSTM [20,21], textCNN [22,23], or BERT [24], and visual features are typically extracted by CNNs. In feature fusion, there are several typical methods as follows. Jin et al. [4] exploit text content, image, and social context to produce a joint representation. An attention mechanism is leveraged to refine the visual representation. However, the refined visual representation cannot reflect the similarity between the visual representation and the social-textual representation, since the attention values are only calculated from the social-textual representation. Wang et al. [5] are inspired by the idea of adversarial networks and thus propose an event adversarial neural network (EANN), which contains an event discriminator used to identify the event label of news, in addition to the feature extractors and the detector. To learn a more general joint representation, a minimax game is set up between the event discriminator and feature extractors. Khattar et al. [6] proposed a multimodal variational autoencoder (MVAE) for fake news detection, which is composed of an encoder, a decoder, and a fake news detector. The encoder first extracts textual features and visual features, which are converted to a sampled multimodal representation. Then, the decoder reconstructs the textual features and visual features from the sampled multimodal representation. Finally, the encoder, the decoder, and the detector are jointly trained to learn a shared representation of multimodal information. Nevertheless, the above three methods [4–6] obtain a joint representation by simply concatenating unimodal features without considering the dependencies between modalities. Song et al. [7] leverage an attention mechanism to fuse a number of word embeddings and one image embedding to obtain fused features, and further extract key features from the fuse features as a joint representation. Although the fusion considers inter-modality relations, it is not fine-grained enough.

2.3. Scaled-Dot Product Attention

The scaled dot-product attention first appears in transformer [9], which is originally used for machine translation tasks. The scaled dot-product attention enables the transformer to capture global dependencies between input and output, which represent text content in two different languages, respectively.

For NLP, Transformer architecture based on the scaled dot-product attention has become the de-facto standard [25]. Some pretrained language models, such as BERT [24], XLNET [26], and GPT-3 [27], have achieved state-of-the-art results on different NLP tasks. Inspired by NLP success, there are multiple works [28,29] that combine CNNs and the scaled dot-product attention in CV. For capturing global information, the scaled dot-product attention has some advantages over repeated convolutional operations, leading to application of the scaled dot-product attention in CV. Thus, some works [25,30] interpret an image as a sequence of words and process them by the Transformer′s encoder solely based on the scaled dot-product attention.

Considering the power of the scaled dot-product attention, we propose to fuse textual features and visual features with the scaled dot-product attention. Like the transformer, the feature fusion in our method is entirely based on the scaled dot-product attention, and the proposed method is expected to improve the performance of fake news detection.

3. Model

3.1. Model Overview

Given news with text and image, the proposed model aims to determine whether the news is real or fake. The architecture of the model is shown in Figure 1, which consists of three parts. The first part is composed of a textual feature extractor and a visual feature extractor, which extract textual features and visual features, respectively. This is followed by the feature fusion, where scaled dot-product attention is used for fine-grained fusion of the textual features and the visual features. The last part is a fake news detector that exploits the fused feature to judge the truth of the news.

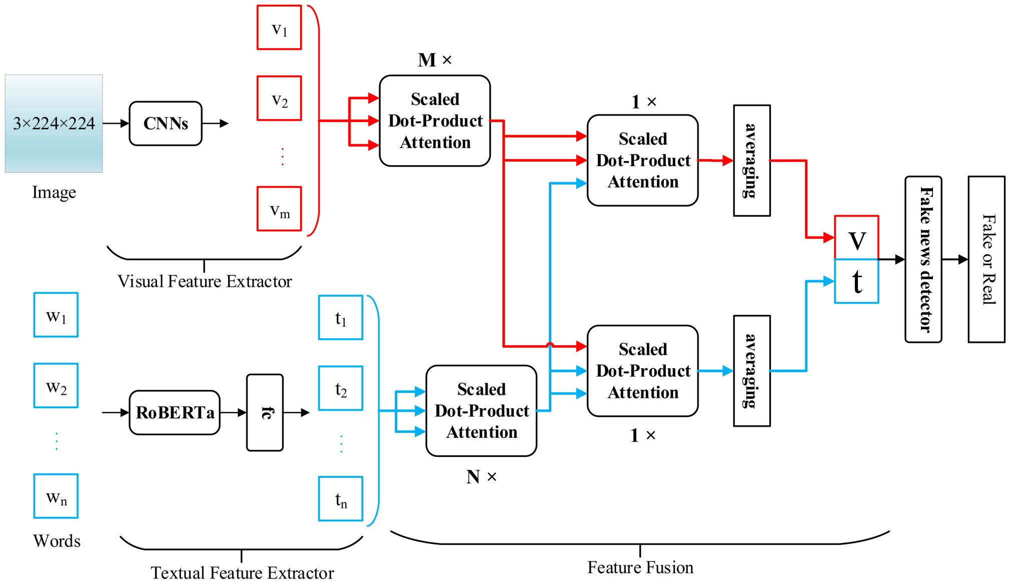

Figure 1. The architecture of our FMFN model.

3.2. Visual Feature Extraction

CNNs have achieved great success in CV. In CNNs, multiple feature maps are obtained by applying convolutional operations of different convolution kernels over an image and can be considered as visual features of the image.

Instead of a visual representation that represents the image, we exploit multiple visual features of the image to fully fuse with textual features, where each visual feature is represented by a feature vector. To learn different features of the image, the VGG-19 [31] is employed, which contains 16 convolutional layers, and 3 feed-forward layers. For an image, the VGG-19 network outputs one vector containing different features, which is not conducive to fine-grained fusion with textual features. Thus, the last three fully-connected layers are removed, and several additional convolutional layers are added behind the 16 convolutional layers of the VGG-19. In this way, the visual feature extractor is composed entirely of convolutional layers and yields a specified number of feature maps $P = [p_1, p_2, \ldots, p_m]$, where m is determined by the number of convolution kernels in the last convolutional layer and each feature map p_i is a $h \times w$ dimensional vector. By collapsing the spatial dimensions of each feature map p_i, we obtain the visual features $R_V = [v_1, v_2, \ldots, v_m]$, each of which is a $hw \times 1$ dimensional vector.

3.3. Textual Feature Extraction

The text content is tokenized into a sequence of tokens denoted as $W = [w_1, w_2, \ldots, w_n]$, where n is the number of tokens. For fine-grained fusion, we obtain the word embedding of each token, rather than a vector representation that represents the text content.

In the NLP field, pretrained language models have achieved state-of-the-art results on different NLP tasks. In particular, the BERT and its variants are widely used due to the ability to utilize both left-to-right and right-to-left contextual information. RoBERTa [8], an improved pretraining procedure for BERT, performs better than BERT on some benchmarks, which removes the next sentence prediction task and adopts the dynamic masking scheme. Thus, RoBERTa is employed to extract word embeddings of the tokens, which is denoted as $E = [e_1, e_2, \ldots, e_n]$.

Compared with other methods of learning word representations, such as word2vec [32], GloVe [33], and fastText [34], word representations generated by the RoBERTa contain contextual information, which means that each word embedding e_i contains information about the entire text content, and therefore can be considered as a textual feature. To adjust the dimensionality of each textual feature, a fully connected layer with ReLU activation function (denoted as "fc" in Figure 1) transforms $E = [e_1, e_2, \ldots, e_n]$ to $R_T = [t_1, t_2, \ldots, t_n]$, where each textual feature t_i is a $d \times 1$ dimensional vector.

3.4. Feature Fusion

Transformer is originally used for machine translation tasks. For a task to translate from English to French, the transformer draws dependencies between English sentences and French sentences thanks to the scaled dot-product attention. We apply the scaled dot-product attention to multimodal fusion so as to capture dependencies between textual features and visual features. In addition, the scaled dot-product attention also can be used to capture global information between these visual features since we extract multiple visual features instead of a visual representation.

Motivated by the above observations, scaled dot-product attention (See Figure 2) is used for fine-grained fusion of textual features and visual features. The scaled dot-product attention block is defined as $ScaledDotProductAttn(Queries, Keys, Values)$, where $Queries$, $Keys$ and $Values$ are mapped into three representations Q, K, and V with three linear layers, then the scaled dot-product attention is computed on Q, K, and V.

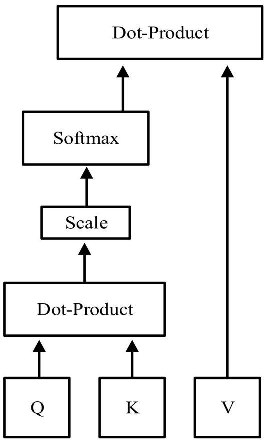

Figure 2. The scaled dot-product attention.

We first enhance the visual features and the textual features using scaled dot-product attention blocks, which can capture global information. For visual features, it enables these features to be further correlated, although global features are obtained by deep CNNs. The process is as follows.

$$R_V^1 = ScaledDotProductAttn(R_V,\ R_V,\ R_V) \tag{1}$$

$$R_V^2 = ScaledDotProductAttn\left(R_V^1,\ R_V^1,\ R_V^1\right) \tag{2}$$

$$R_V^M = ScaledDotProductAttn\left(R_V^{M-1},\ R_V^{M-1},\ R_V^{M-1}\right) \tag{3}$$

where M is the number of the scaled dot-product attention blocks and $R_V^M = \left[v_1^M,\ v_2^M,\ \ldots,\ v_m^M\right]$ represents a number of enhanced visual features. Several scaled dot-product attention blocks (The number of the blocks is N) are also applied to the textual features R_T to obtain $R_T^N = \left[t_1^N,\ t_2^N,\ \ldots,\ t_n^N\right]$ in the same way.

Then, two scaled dot-product attention blocks are utilized to refine the enhanced visual features R_V^M and the enhanced textual features R_T^N, respectively. The process to refine the visual features R_V^M is as follows.

$$R'_V = ScaledDotProductAttn\left(R_T^N,\ R_V^M,\ R_V^M\right) \tag{4}$$

The $R'_V = \left[v'_1,\ v'_2,\ \ldots,\ v'_m\right]$ are the refined visual features representing the fine-grained fusion with the textual features R_T^N. Note that the queries come from the enhanced textual features, and the keys and the values come from the enhanced visual features. Therefore, it can capture the dependencies between visual features and textual features. The R'_T is also obtained by computing the scaled dot-product attention, where queries come from the enhanced visual features, and the keys and the values come from the enhanced textual features.

Finally, the refined features R'_V and R'_T are transformed to two vectors v and t by the averaging. The process of averaging the refined features R'_V to produce the vector v is as follows.

$$v = \frac{v'_1 \oplus v'_2 \oplus \ldots \oplus v'_m}{m} \tag{5}$$

where $\oplus$ denotes element-wise sum. The two vectors v and t are concatenated into a vector r as the joint representation, which not only considers the correlations between different visual features but also reflects the dependencies between textual features and visual features.

3.5. Fake News Detector and Model Learning

The fake news detector is a fully connected layer with SoftMax function, which takes the joint representation r as input to make the prediction as follows.

$$\hat{y} = softmax(W \times r + b) \tag{6}$$

where W is parameters of the fully connected layer and b is the bias term.

To configure the model for training, the loss function is set to cross entropy as follows.

$$L(\theta)\ = -ylog(\hat{y}) - (1-y)log(1-\hat{y}) \tag{7}$$

where θ represents all of the learnable parameters of the proposed model, and $y \in \{0,\ 1\}$ denotes the ground-truth label.

4. Experiments

4.1. Dataset

We evaluate the effectiveness of the proposed model on the dataset collected by Jin et al. [4], on which the real news is collected from an authoritative news source, Xinhua News Agency, and the fake news is verified by Weibo’s official rumor debunking system.

For the dataset, we only focus on tweets with text and images in order to fuse textual features and visual features. Thus, tweets without text or images are removed. The data split scheme is the same as the benchmark scheme, and the data are preprocessed in a similar way to the work [4]. The detailed statistics of the dataset are listed in Table 1.

Table 1. The Weibo dataset.

	Training Set	Test Set
fake news	3345	862
real news	2807	835
images	6152	1697

4.2. Settings

The optimizer used is Adam [35] with a learning rate of 0.001, $\beta 1 = 0.9$ and $\beta 2 = 0.999$.

For the textual feature extraction, the Chinese BERT with whole word masking [36,37] is used, and the max length of text is set to 160. For efficient training, the feature-based approach is adopted on the pretrained language model, which means that the parameters of the pretrained language model are fixed. Only the fully connected layer with ReLU activation function (denoted as "fc" in Figure 1) is trained, and its hidden size is 100.

For the visual feature extraction, the first 16 convolutional layers and the first four max-pooling layers of VGG19 are adopted, which means that we remove the last three fully-connected layers, and the last max-pooling layer of VGG19. The parameters of the 16 convolutional layers are frozen. Two additional convolutional layers with ReLU activation function, the first with 256 convolution kernels and the second with 160 convolution kernels, are added behind these layers and trained. For these convolution kernels, the receptive field is 3×3, and the convolution stride is 1. Thus, 160 visual features are produced by the visual extractor, each of which is a 100×1 dimensional vector.

As above, the number of visual features m is equal to the number of textual features n, and the dimensionality of each visual feature and each text feature are also equal, which facilitates the computation of the Scale-Dot Product Attention.

For the M and N, they are set to 3 and 1, respectively, which achieves the best performance.

4.3. Baselines

For comparison with other methods, two unimodal models and six multimodal models are chosen as baselines, which are listed as follows:

- Textual: All scaled dot-product attention blocks and the visual feature extractor are removed from the proposed model FMFN. The textual features R_T obtained by the textual feature extractor are transformed to a vector by the averaging, and the vector is fed into a binary classifier to train a model. For a fair comparison, the parameters of the RoBERTa in the textual feature extractor are frozen.
- Visual: Similar to textual, the visual feature extractor, and a binary classifier are jointly trained for fake news detection. For a fair comparison, the parameters of the first 16 convolutional layers in the visual feature extractor are fixed.
- VQA [38]: The objective of visual question answering is to answer questions concerning certain images. The multi-class classifier in the VQA model is replaced with a binary classifier, and one-layer LSTM is used for a fair comparison.
- NeuralTalk [39]: The model aims to produce captions for given images. The joint representation is obtained by averaging the outputs of RNN at each time step.
- att-RNN [4]: A novel RNN with an attention mechanism is utilized to fuse multimodal features for effective rumor detection. For a fair comparison, we do not consider the social context, and only fuse textual features and visual features.
- EANN [5]: The model is based on adversarial networks, which can learn event-invariant features containing multimodal information.

- MVAE [6]: By jointly training the VAE and a classifier, the model is able to learn a shared representation of multimodal information.
- CARMN [7]: An attention mechanism is used to fuse word embeddings and one image embedding to obtain fused features. From the fuse features, key features are extracted as a joint representation.

4.4. Comparison with Baselines

Table 2 shows the results of different methods on Weibo dataset. We can observe that our proposed model achieves competitive results.

Table 2. The results of different methods on Weibo dataset.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
Textual	0.725	0.763	0.661	0.708	0.677	0.774	0.722
Visual	0.657	0.682	0.617	0.648	0.622	0.68	0.65
VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.76
NeuralTalk	0.726	0.794	0.613	0.692	0.684	0.84	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
CARMN	0.853	**0.891**	0.814	0.851	0.818	0.894	0.854
FMFN	**0.885**	0.878	**0.851**	**0.864**	**0.874**	**0.896**	**0.885**

Specifically, the proposed model FMFN achieves an accuracy of 88.5% on the dataset and outperforms all of the baseline models except the precision of fake news. In these baseline systems, CARMN performs best, which can be attributed to the attention mechanism. The attention mechanism in CARMN can capture the dependencies between textual features and visual features, but other multimodal methods, which simply concatenate unimodal features, cannot learn the dependencies. The dependencies include consistency between text content and image content. The news with inconsistent text and image is generally fake. It is difficult to identify if the dependencies between textual features and visual features cannot be captured. Compared with CARMN, our model boosts accuracy by about 3%. It is the fined-grained fusion of word embeddings and multiple visual features that achieves significant improvements, whereas CARMN only fuses word embeddings and one image embedding. It illustrates the importance of the fine-grained fusion, which facilitates a better capture of such dependencies.

5. Ablation Analysis

5.1. Component Analysis

To verify the impact of each component of FMFN, three baselines are constructed as follows.

- FMFN(CONCAT): The last two scaled dot-product attention blocks are removed from the proposed model FMFN. By the averaging, the R_V^M and R_T^N are transformed to two vectors, respectively. The concatenation of the two vector is fed into the fake news detector. Therefore, it cannot capture the dependencies between textual features and visual features.
- FMFN(TEXT): We do not use the refined visual features R'_V and only use the refined textual features R'_T. The refined textual features R'_T are transformed to a vector by the averaging, and the vector is fed into the fake news detector.
- FMFN (M = 0): The number of scaled dot-product attention blocks M is set to 0, which means that we do not consider the correlations between different visual features.

From Table 3, we can see that our proposed method FMFN outperforms all baselines. If we remove one of the components from the model, both the accuracy and F_1 scores will drop. The results show that all components of the model are indispensable.

Table 3. The results of FMFN (CONCAT), FMFN (TEXT), FMFN (M = 0), and FMFN.

Method	Accuracy	Fake News F_1	Real News F_1
FMFN (CONCAT)	0.867	0.839	0.872
FMFN (TEXT)	0.874	0.845	0.876
FMFN (M = 0)	0.877	0.851	0.880
FMFN	**0.885**	**0.864**	**0.885**

Compared with FMFN (CONCAT), the accuracy of FMFN increases from 86.7% to 88.5%. It shows that the scaled dot-product attention blocks used to capture the dependencies between visual features and textual features are critical for performance improvement. For FMFN (CONCAT), simply concatenating multiple visual features and textual features can yield relatively good results (an accuracy of 86.7%) without using attention, which shows the importance of representing different features of an image with multiple feature vectors. If we only use the refined textual features, the accuracy will drop about 1%, which indicates that both the refined textual features and the refined visual features are important. For the hyper-parameter M, there will be a performance loss as well if we set it to 0. This indicates that it is useful to use attention to make multiple visual features correlated.

5.2. Visualization of the Joint Representation

To further illustrate the impact of the feature fusion, the joint representation r learned by FMFN and the joint representation learned by FMFN(CONCAT) are visualized with t-SNE [40]. As depicted in Figure 3, two colors represent fake news and real news, respectively.

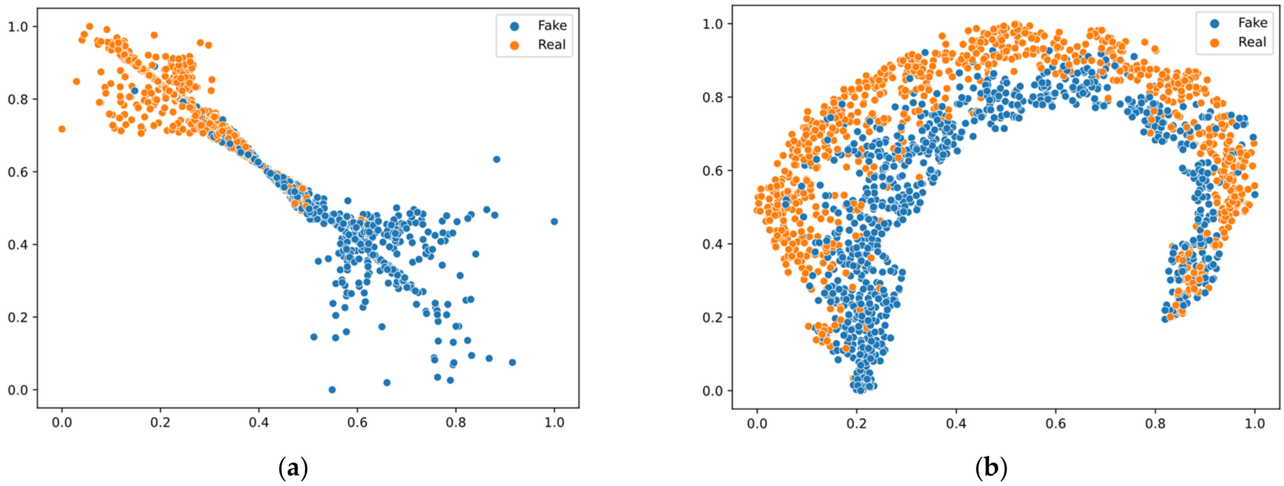

Figure 3. Visualization of the joint representation: (**a**) FMFN; (**b**) FMFN (CONCAT).

From Figure 3, we can see that FMFN can learn more discriminable representations compared with FMFN (CONCAT). As is shown in Figure 3a, the representations of the different categories are in the upper left and lower right regions of the image. In addition, the representations of the same category are more easily aggregated, which makes the number of points in Figure 3a look small. For FMFN (CONCAT), it basically distinguishes between two types of representations. However, there are many representations that are difficult to distinguish. The visualization illustrates the effectiveness of the feature fusion.

6. Conclusions and Future Work

We propose a novel fine-grained multimodal fusion network (FMFN) to fully fuse textual features and visual features for fake news detection. For a tweet with text and image, multiple different visual features of the image are obtained by deep CNNs and word embeddings of words in the text are extracted by a pretrained language model, each of which can be considered as a textual feature. The scaled dot-product attention is employed to enhance the visual features as well as the textual features and fuse them. This is a fine-grained and adequate fusion, which not only considers the correlations between different visual features but also captures the dependencies between textual features and visual features. Experiments conducted on a public Weibo dataset demonstrate the effectiveness of FMFN. In comparison with other methods for fusing the visual representation and the text representation, FMFN achieves competitive results. It shows that the joint representation learned by the FMFN, which fuses multiple visual features and multiple textual features, is better than the joint representation obtained by fusing a visual representation and a text representation in determining fake news.

In the future, we plan to fuse social context features in addition to textual features and visual features. Moreover, the visual features in the frequency domain [17] are considered to further improve the performance of fake news detection.

Author Contributions: Conceptualization, H.L. and H.M.; investigation, J.W.; methodology, H.L., H.M. and J.W.; software, H.L.; project administration, J.W.; validation, H.L., H.M. and J.W.; writing—original draft preparation, H.L.; writing—review and editing, H.L., H.M. and J.W.; visualization, J.W., H.L. and H.M.; funding acquisition, J.W. and H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 61872145, 62032024), MOE International Joint Lab of Trustworthy Software, East China Normal University (No. 2021-5) and the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University (No. ESSCKF2021-03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Czeglédi, C.; Valentinyi, K.V.; Borsos, E.; Járási, É.; Szira, Z.; Varga, E. News Consuming Habits of Young Social Media Users in the Era of Fake News. *WSEAS Trans. Comput.* **2019**, *18*, 264–273.
2. Helmstetter, S.; Paulheim, H. Collecting a Large Scale Dataset for Classifying Fake News Tweets Using Weak Supervision. *Future Internet* **2021**, *13*, 114. [CrossRef]
3. Zakharchenko, A.; Peráček, T.; Fedushko, S.; Syerov, Y.; Trach, O. When Fact-Checking and 'BBC Standards' Are Helpless: 'Fake Newsworthy Event' Manipulation and the Reaction of the 'High-Quality Media' on It. *Sustainability* **2021**, *13*, 573. [CrossRef]
4. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 795–816.
5. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. 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, London, UK, 19–23 August 2018; pp. 849–857.
6. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
7. Song, C.; Ning, N.; Zhang, Y.; Wu, B. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Inf. Process. Manag.* **2021**, *58*, 102437. [CrossRef]
8. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.

9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
10. Islam, N.; Shaikh, A.; Qaiser, A.; Asiri, Y.; Almakdi, S.; Sulaiman, A.; Moazzam, V.; Babar, S.A. Ternion: An Autonomous Model for Fake News Detection. *Appl. Sci.* **2021**, *11*, 9292. [CrossRef]
11. Alonso, M.A.; Vilares, D.; Gómez-Rodríguez, C.; Vilares, J. Sentiment Analysis for Fake News Detection. *Electronics* **2021**, *10*, 1348. [CrossRef]
12. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.-F.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 3818–3824.
13. Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. A convolutional approach for misinformation identification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3901–3907.
14. Ma, J.; Gao, W.; Wong, K.-F. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 3049–3055.
15. Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; Tian, Q. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Trans. Multimed.* **2017**, *19*, 598–608. [CrossRef]
16. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
17. Qi, P.; Cao, J.; Yang, T.; Guo, J.; Li, J. Exploiting Multi-domain Visual Information for Fake News Detection. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 518–527.
18. Wu, K.; Yang, S.; Zhu, K.Q. False rumors detection on Sina Weibo by propagation structures. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 13–17 April 2015; pp. 651–662.
19. Liu, Y.; Wu, Y.-F.B. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018.
20. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal. Process.* **1997**, *45*, 2673–2681. [CrossRef]
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014. [CrossRef]
23. Zhang, Y.; Wallace, B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv* **2015**, arXiv:1510.03820.
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
26. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 5753–5763.
27. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
28. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. *End-to-End Object Detection with Transformers*; Springer: Cham, Switzerland, 2020; pp. 213–229.
30. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R.B. Masked Autoencoders Are Scalable Vision Learners. *arXiv* **2021**, arXiv:2111.06377.
31. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
32. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Harrahs and Harveys, Lake Tahoe, NV, USA, 5–10 December 2013.
33. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
34. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv* **2016**, arXiv:1607.01759.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980.
36. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv* **2019**, arXiv:1906.08101. [CrossRef]
37. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. *arXiv* **2020**, arXiv:2004.13922.

38. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
39. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
40. Van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.