Similarity-Aware Attention Network for Multimodal Fake News Detection

Diwen Dong, Fuqiang Lin, Guowei Li and Bo Liu

National University of Defense Technology, Changsha, China

Abstract

The wide spread of online fake news has drawn a growing concern since its damage to public trust. Images play an important role in detecting fake news as part of the posts on social media. Previous works have made achievements by focusing on either the complementary information of the image-text pair or the cross-modal inconsistency. However, few pieces of research focus on leveraging both types of information in a unified framework. Besides, due to the intrinsic gaps between the text and the image, the inconsistent information could be difficult to capture. In this paper, we propose a Similarity-Aware Attention Network (SAAN), a multimodal fake news detection method with an attention-based feature extractor to capture the textual feature, visual feature, and cross-modal complementary information sufficiently and flexibly, as well as a CLIP-guided similarity evaluator to measure the inconsistency between the text and image in the same semantic space. We also design a similarity-based loss to benefit fake news prediction by increasing the gap between fake news and real news in representation. Experiments on two real-world datasets indicate the superiority of our proposed SAAN and the effectiveness of the designed modules.

Keywords

Fake news, multimodal learning, neural networks

1. Introduction

Online dissemination of fake news has become a severe problem for the public. Fake news in a broad definition[1] contains all types of false information published on social media such as Twitter and Weibo, which can mislead people, trigger panic, and damage public trust in government. It even has the power to influence the 2016 U.S. presidential election [2]. The low cost of manufacture and high speed of spread makes it difficult to detect fake news manually. Therefore, automatic fake news detection has become a growing concern. Some previous works about fake news detection have focused on text modality and proposed some methods such as writing style-based [3], statistics-based [4] and deep neural models with textual features [5, 6].

However, detecting fake news with only text modality is not complete and sufficient. First, much news is posted on social media with one or more images, which contain much semantic information. Second, research [7] has indicated that the characteristics of the image itself can provide clues, such as traces of tampering, for fake news detection. As an approach to improve the performance of the classifier, several works take visual information into consideration and propose a series of methods for multimodal fake news detection. In addition to fusing textual and visual features with concatenation [8], here are two types of information mainly used in the previous works: (1) complementary information and (2) inconsistent information. On the one hand, the text and image constituting whole news are generally associated with and enhance each other semantically. Series of methods [9, 10] have been proposed to capture the complementary information. On the other hand, it is hard to find a perfectly

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

EMAIL: ddw_bak@nudt.edu.cn (Diwen Dong); linfuqiang13@nudt.edu.cn (Fuqiang Lin); liguowei@nudt.edu.cn (Guowei Li); kyle.liu@nudt.edu.cn (Bo Liu)

ORCID: 0000-0002-6364-9662 (Diwen Dong); 0000-0001-9314-9493 (Fuqiang Lin); 0000-0003-1801-207X (Guowei Li); 0000-0002-9953-8438 (Bo Liu)

matching image for the fabricated article, thus making inconsistency of image-text pair a common phenomenon in fake news. Zhou et al. [11] design a similarity-based loss to capture the cross-modal inconsistency between the text and image.

Although previous works have achieved promising results, there are still some issues to be optimized for multimodal fake news detection. First, there are significant gaps between text and image, thus making the cross-modal similarity in- appropriate. For example, [11] projects image to text by a visual caption model, which has limitations in mapping text and image to the same semantic space and introduces noise to the similarity calculation. Second, the information density of features from different modalities is distinct, so the depth of encoding before fusion ought to differ for fully capturing the complementary information. Third, there are not many multimodal methods combining both complementary and inconsistent information. The two types of information exhibit distinct effectiveness in different circumstances, so finding an available way to utilize them together is critical.

In this paper, we propose a Similarity-Aware Attention Network (SAAN) for multimodal fake news detection. Specifically, we design a flexible attention-based multimodal feature extractor, which consists of a text/image encoder to get the global and local embeddings of text and image, a self-attention-based unimodal feature encoding module to obtain high-quality feature representations, and a co-attention-based multimodal feature fusion module to fully capture the correla0tion between features from different modalities. In addition, we leverage a Contrastive Language-Image Pre-training (CLIP) model to project the text and image to the same semantic space to reduce the gaps between them and design a similarity-based loss as an auxiliary to improve the performance of the fake news detection model. The contributions of this paper can be summarized as follows:

- We propose SAAN, a multimodal fake news detection method aggregating both the complementary and inconsistent information of news posts.
- We design an attention-based feature extractor to capture the textual feature, visual feature, and cross-modal complementary information sufficiently and flexibly. Besides, we design a CLIP-guided similarity evaluator to measure the inconsistency between the text and image in the same semantic space.
- We have conducted comprehensive experiments on two real-world datasets, and our proposed model overperforms all the baselines. The results of the ablation study indicate the effectiveness of independent components of SAAN.

2. Related Work

2.1. Unimodal Fake News Detection

Unimodal fake news detection focuses on extracting features of either the text or image of the news post.

For texts, early works using handcrafted features tend to concentrate on statistics of articles [4], mismatched headlines [12] and writing style [3]. With the development of deep learning, recent researchers leverage deep neural networks [5, 6] to learn the representation of text. Chen et al. [5] propose a CNN combined with an attention-residual network for fake news detection based on the text of the post. Vaibhav et al. [6] propose a graph neural network-based model which breaks away from the need for feature engineering to fine-grained fake news classification.

For images, Cao et al. [7] explore multiple visual characteristics for fake news detection, including semantic features, forensics features, context features, and statistical features. Experimental results show that detecting the traces of tampering in images is beneficial to fake news detection. In addition, the quality of images [13], as well as inconsistency between visual entities and external knowledges [14], could also help to the prediction.

2.2. Multimodal Fake News Detection

Most news on social media is composed of a post with one or more images attached. Recently, many researchers have concentrated on the importance of images for fake news detection. Singhal et al. [8] propose a multimodal framework with a text encoder and an image encoder to extract different kinds of

features, which provides a basic pattern for multimodal fake news detection with deep learning. Chen et al. [9] utilize the self-attention mechanism to fuse textual and visual features and introduces a latent topic memory module to store the semantic information about real and fake news events. Wu et al. [10] design a cross-modal attention fusion mechanism to capture the latent correlations of text and image and leverage a Bi-GRU to extract sequential information of text properly. In addition to the cross-modal complementary information, some works focus on the inconsistency between text and image. Zhou et al. [11] design a new loss from the perspective of measuring the mismatches between news content and the attached image. Due to the previous achievements, not many works consider combining both complementary information and conflicting information between modalities.

3. Method

3.1. Overview

Figure 1 shows the architecture of our proposed SAAN model. It consists of three main components: (1) Attention- based multimodal feature extractor, (2) CLIP-guided similarity evaluator, and (3) Fake news predictor. Attention-based multi- modal feature extractor is a component of a text/image encoder to obtain the original unimodal features, a self-attention-based encoder to get deeper representations of text and image, and a co-attention module to fully extract cross-modal complementary information. The input text and image are also fed to the CLIP-guided similarity evaluator, which is designed to calculate the inter-modal similarity and fine-tuned to lower the similarity-based loss. Finally, we concatenate the output feature of the attention-based multimodal feature extractor and CLIP-guided similarity evaluator. The combined feature is sent to the fake news predictor to calculate the binary-cross-entropy-based loss. Two types of loss are optimized together with the ground-truth label in the training stage.

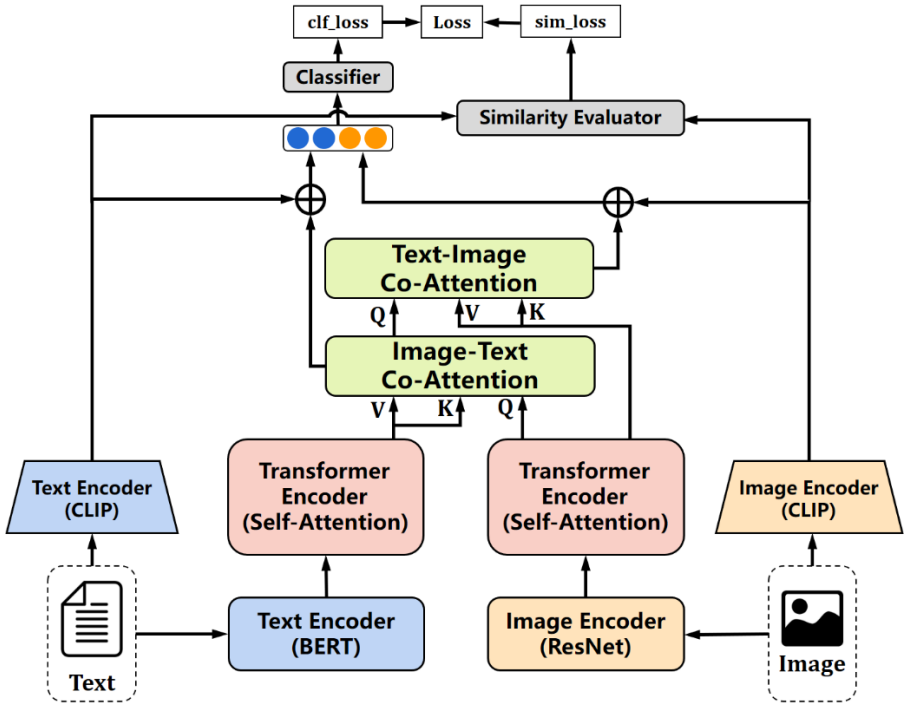

Figure 1: The architecture of SAAN.

3.2. Attention-Based Multimodal Feature Extractor

3.2.1. Text Encoder

Given a sequence of input text $\mathcal{T}$, we employ a pre-trained BERT [15] to obtain the textual representation $\boldsymbol{R}^{\mathcal{T}}_{\text{BERT}}$. We first input the raw text to the BERT tokenizer, which adds a [CLS] token at

the beginning of the text and then tokenizes sentences to a sequence of tokens. The length of the sequence is limited to 1. The process can be denoted as

$$\boldsymbol{R}_{\mathrm{BERT}}^{\mathcal{T}} = \{t_0, t_1, t_2, \dots, t_l\} = \mathrm{BERT}([\mathrm{CLS}], w_1, w_2, \dots, w_m), \tag{1}$$

where m is the original length of $\mathcal{T}$ and t_i the i-th text token. $\boldsymbol{R}_{\mathrm{BERT}}^{\mathcal{T}} \in \mathbb{R}^{l \times d}$ where d is the last hidden layer dimension of BERT.

3.2.2. Image Encoder

For an input image $\mathcal{V}$, we first leverage a pre-trained Faster R-CNN model for object detection. After that, $\mathcal{V}$ is split into several visual regions. Then a pretrained ResNet50 is utilized to obtain the visual representation $\boldsymbol{R}_{\mathrm{ResNet}}^{\mathcal{V}}$. We encode the whole image and visual regions with ResNet50 as global and local features to capture multi-scale visual features and align with the textual representation. The output representation of the vision model ResNet is given by:

$$\boldsymbol{R}_{\mathrm{ResNet}}^{\mathcal{V}} = \{v_0, v_1, v_2, \dots, v_l\} = \{\mathrm{MP}(\mathrm{ResNet}(b_i))\} \,|i \in [0, n]\,, \tag{2}$$

where b_i is the i-th region of $\mathcal{V}$, b_0 represents the whole image and n is the number of all detected regions. To match the attributes of the textual representation, we limit the length of $\boldsymbol{R}_{\mathrm{ResNet}}^{\mathcal{V}}$ to l and resize the dimension of each visual vector v_i to d by an adaptive Mean Pooling (MP) operation.

3.2.3. Unimodal Feature Encoding

The Unimodal Feature Encoding module aims to produce deeper news content representation $\mathbf{R}_{\mathrm{Uni}}^{\mathcal{T}}$ and news image representation $\boldsymbol{R}_{\mathrm{Uni}}^{\mathcal{V}}$. To capture high-quality text and image features, we leverage Transformer Encoder [16] which is based on the self-attention mechanism, as the module's core. Setting the number of self-attention layers in a Transformer Encoder is flexible so that we can build the module according to the information density of features from different modalities. Specifically, for an input feature vector $\boldsymbol{R}_{in}$, the output $\boldsymbol{R}_{out}$ from 1-layer Transformer Encoder is calculated as follows:

$$\widetilde{\boldsymbol{R}} = \mathrm{MultiHeadAttention}(\boldsymbol{R}_{in}), \tag{3}$$

$$\boldsymbol{R} = \mathrm{LayerNorm}\big(\widetilde{\boldsymbol{R}} + \boldsymbol{R}_{in}\big), \tag{4}$$

$$\boldsymbol{R}' = \mathrm{FeedForwardNetwork}(\boldsymbol{R}), \tag{5}$$

$$\boldsymbol{R}_{out} = \mathrm{LayerNorm}(\boldsymbol{R}' + \boldsymbol{R}), \tag{6}$$

where $\tilde{R}$, R, and R' are intermediate results.

For textual feature, $\boldsymbol{R}_{in}$ represents $\boldsymbol{R}_{\mathrm{BERT}}^{\mathcal{T}}$. For visual feature, $\boldsymbol{R}_{in}$ represents $\boldsymbol{R}_{\mathrm{ResNet}}^{\mathcal{V}}$. The global and local features could be fully merged in the above process. Since the semantic information in the text is more prosperous than in image generally, we employ a 2-layer Transformer Encoder for the textual feature and a 1-layer Transformer Encoder for the visual feature.

3.2.4. Multimodal Feature Encoding

To characterize the relative importance of regions and tokens, we design two attention networks based on the co-attention mechanism, named image-text attention and text-image attention. The former allows the model to consider the contribution of different visual regions to text tokens, while the latter captures the importance of different tokens to visual regions. The calculation of the attention is formulated as follows:

$$\mathrm{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{7}$$

$$Multihead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat(\mathrm{Attn}_1, \mathrm{Attn}_2, \ldots, \mathrm{Attn}_\mathrm{H}), \tag{8}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the matrices to obtain queries, keys, and values, d_k is the dimension of queries and keys, and H represents the number of heads.

As shown in Figure 1, the queries and keys are calculated by visual representation, and the values are obtained from textual representation in image-text attention. Correspondingly, in text-image attention, the queries and keys come from text features, and the values come from image features to measure the importance of each token to all the visual regions. Each region/token is assigned a weight α to denote its attribution via calculating the cosine similarity between tokens and regions:

$$\boldsymbol{Q} = \boldsymbol{W_Q}\boldsymbol{R}_{co}, \tag{9}$$

$$\boldsymbol{K} = \boldsymbol{W_K}\boldsymbol{R}_{in}, \tag{10}$$

$$\boldsymbol{V} = \boldsymbol{W_V}\boldsymbol{R}_{in}, \tag{11}$$

where $\boldsymbol{R}_{in}$ represents $\boldsymbol{R}^{\mathcal{V}}_{\mathrm{Uni}}$ in image-text attention and $\mathbf{R}^{\mathcal{T}}_{\mathrm{Uni}}$ in text-image attention, $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, and $\boldsymbol{W_V}$ are trainable metrics. We connect the two modules in series to obtain the new representations of the text and image, denoted as $\boldsymbol{R}^{\mathcal{T}}_{\mathrm{Multi}}$ and $\boldsymbol{R}^{\mathcal{V}}_{\mathrm{Multi}}$, respectively.

3.3. CLIP-Guided Similarity Evaluator

Though the inner and inter modalities information is extracted by the above networks, semantic gaps remain between the text and image features. Therefore, it is significant to project the text and image to a common semantic space to effectively evaluate the inconsistency between modalities.

Inspired by the previous work [11], we design a CLIP-Guided Similarity Evaluator with a similarity-based loss as an auxiliary to capture the cross-modal inconsistent information. First, we use a CLIP model to map the text and image to the same representative space. CLIP is a multimodal model pretrained on a large amount of image-text pairs, which has a strong ability to learn the intrinsic correlation between text and image. It consists of an image encoder and a text encoder, which we leverage to re-encode the news content and the attached image. We denote the CLIP-encoded text and image features as $\boldsymbol{R}^{\mathcal{T}}_{\mathrm{CLIP}}$ and $\boldsymbol{R}^{\mathcal{V}}_{\mathrm{CLIP}}$. Then we calculate the similarity of the new textual and visual features by:

$$Sim = \frac{\boldsymbol{R}^{\mathcal{T}}_{\mathrm{CLIP}} \cdot \boldsymbol{R}^{\mathcal{V}}_{\mathrm{CLIP}}}{||\boldsymbol{R}^{\mathcal{T}}_{\mathrm{CLIP}}|| \cdot ||\boldsymbol{R}^{\mathcal{V}}_{\mathrm{CLIP}}||}. \tag{12}$$

To guarantee $Sim \in [0,1]$, we apply a Sigmoid function to it:

$$Sim' = \mathrm{Sigmoid}(Sim). \tag{13}$$

3.4. Fake News Prediction

3.4.1. Feature Aggregation

To obtain an integrated presentation of text and image, we merge the features from the attention-based multimodal feature extractor and the CLIP model:

$$\boldsymbol{R}^{\mathcal{T}} = \mathrm{Concat}\left(\boldsymbol{R}^{\mathcal{T}}_{\mathrm{Multi}}, \boldsymbol{R}^{\mathcal{T}}_{\mathrm{CLIP}}\right), \tag{14}$$

$$\boldsymbol{R}^{\mathcal{V}} = \mathrm{Concat}\left(\boldsymbol{R}^{\mathcal{V}}_{\mathrm{Multi}}, \boldsymbol{R}^{\mathcal{V}}_{\mathrm{CLIP}}\right), \tag{15}$$

$$\boldsymbol{R}^{Agg} = \mathrm{Concat}\left(\boldsymbol{R}^{\mathcal{T}}, \boldsymbol{R}^{\mathcal{V}}\right), \tag{16}$$

where Concat refers to the concatenating operation.

3.4.2. Classification and objective function

We design two types of loss for fake news detection: a binary-cross-entropy-based loss and a similarity-based loss. We feed the aggregated feature R^{Agg} to an MLP layer and employ a sigmoid function to obtain the prediction $\hat{y}$. Then the binary-cross-entropy-based loss is calculated as:

$$\mathcal{L}_{clf} = ylog(\hat{y}) + (1-y)log(1-\hat{y}) , \quad (17)$$

where y is the ground-truth label ('fake' maps to 0 and 'real' maps to 1).

Based on the assumption that the probability of mismatches between text and image of fake news is much higher than real news, the similarity-based loss is designed as:

$$\mathcal{L}_{sim} = ylog(Sim') + (1-y)log(1-Sim') . \quad (18)$$

It is worth mentioning that the CLIP model is fine-tuned during training while the parameter of BERT and ResNet are frozen. Finally, we specify the final loss as:

$$\mathcal{L} = \alpha\mathcal{L}_{clf} + \beta\mathcal{L}_{sim} , \quad (19)$$

where α and β are hyperparameters.

4. Experiments

4.1. Datasets

We conduct experiments on two real-world datasets in English and Chinese, relatively named *Twitter* and *Weibo*. The statistics of the two datasets are shown in Table 1. To verify the effectiveness of our proposed method, we filter out samples without text or images.

The *Twitter* dataset was released for the Verifying Multimedia Use Task [17] and widely used in previous works. Following the original partition, we split Twitter into 13062/831 as Train/Test set in experiments for fair competition.

The *Weibo* dataset was collected from Sina Weibo, one of the most effective social media in China. We use a public version released by Jin et al. [18] and split it into 5482/672/1699 as Train/Dev/Test in experiments.

Table 1

The Statistics of *Twitter* and Weibo Datasets.

	Twitter	*Weibo*
# of real news	5,870	3,642
# of fake news	8,023	4,211
# of images	410	7,853

4.2. Implement Details

We use Huggingface pretrained language models *bert-base- uncased*[1] and *bert-base-chinese*[2] as the text encoder for *Twitter* and *Weibo*, relatively. For images, we use the pretrained Faster R-CNN[3] for object detection and ResNet50[4] for encoding visual regions. All regions were shaped to a size of 224×224. The dimension of textual and visual features is 768. In addition, we limit the max length of input sequences to 31. The weights of BERT, Faster R-CNN, and ResNet50 are frozen in the training stage. We leverage the official version of pretrained CLIP named *ViT-B/32*[5] for *Twitter*. For Weibo dataset, we use an open source CLIP model[6] pretrained on chinese corpus. Since the distinction of

[1] https://huggingface.co/bert-base-uncased
[2] https://huggingface.co/bert-base-chinese
[3] https://pytorch.org/vision/stable/models/faster rcnn.html
[4] https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html
[5] https://github.com/openai/CLIP
[6] https://huggingface.co/IDEA-CCNL/Taiyi-CLIP-Roberta-large-326M-Chinese

information density between text and image, we use a 2-layer self-attention module for text and a 1-layer self-attention module for the image. A 2-layer co-attention module is used to capture the cross-modal features. The Adam optimizer [19] is adopted for training, and we set the learning rate as 1e-5. The batch size is set to 32 for *Twitter*, 64 for *Weibo*, and the epoch is set to 100 with an early stopping mechanism to avoid over-fitting. The α and β in 11 are selected as 1.0 and 0.5, respectively.

4.3. Baselines

We compare our proposed model with several existing multimodal approaches for fake news detection to evaluate its effectiveness. The baselines are listed as follows:

- **EANN** [20] is an end-to-end framework with an event discriminator to remove the event-specific features and keep shared features among events, thus benefiting fake news detection.
- **MVAE** [21] trains a variational autoencoder, which is capable of learning shared representations for image and text, thereby discovering correlations between modalities for multimodal fake news detection.
- **SpotFake** [8] uses a VGG-19 as an image encoder to extract the visual features and a pretrained BERT as a text encoder to obtain textual features. The two types of feature vectors are then concatenated to the fake news classifier.
- **SAFE** [11] first extract textual and visual features separately with neural networks and design a loss based on the similarity of the text and image based on the assumption that fake news tends to use irrelevant images.
- **MFN** [5] utilizes the self-attention mechanism to fuse textual and visual features and introduces a latent topic memory module to store the semantic information about real and fake news events.
- **CALM** [10] designs a cross-modal attention fusion mechanism to capture the latent correlations of text and image and leverage a Bi-GRU to extract sequential information of text properly.
- **CAFE** [22] proposes an ambiguity-aware multimodal fake news detection method with a cross-modal ambiguity learning module to estimate the ambiguity between different modalities and a cross-modal fusion module to capture the cross-modal correlations.

4.4. Main Results

Table 2 and Table 3 show the performance of our proposed SAAN on *Twitter* and *Weibo*, respectively.

First, we can find that our proposed SAAN achieves better performance than all the baselines on the two datasets. Specifically, our method outperforms 8% in accuracy and 12.1% in F1 score than CALM on Twitter. On Weibo dataset, it gains an improvement of 0.7% in accuracy and 0.1% in F1 score, inferior to the performance on English datasets. One of the possible reasons is the reduction in the Chinese pretraining corpus of the CLIP model, causing a decline in measuring the similarity between text and image. For other metrics, SAAN also shows superiority among compared methods, demonstrating the effectiveness of our method in the fake news detection task.

Besides, the contrast among different kinds of methods shows the significance of the fusion manner to the final performance. Methods with a fused feature vector obtained by simply concatenating text and image features, such as EANN and SpotFake, lack sufficient cross-modal correlation information and ignore the inconsistency between textual and visual information. Thus, their performance is lower than approaches that concentrate more on multimodal fusion. SAFE leverages the inconsistency by evaluating the mismatches between two types of features, but the image-to-text model has a limited ability to project images to the same semantic space as texts. Our proposed SAAN adopts a cross-modal co- attention module to extract the complementary information between modalities and a CLIP-guided similarity evaluator to evaluate the contradiction between text and image, boosting the performance of the fake news classifier.

Table 2
Results of Comparison among Different Models on *Twitter*.

Method	Acc.	Prec.	Recall	F_1
EANN	0.715	0.822	0.638	0.719
MVAE	0.805	0.869	0.588	0.702
SpotFake	0.778	0.751	0.900	0.820
MFN	0.806	0.799	0.777	0.785
CALM	0.845	0.785	0.831	0.807
SAAN	**0.925**	**0.915**	**0.941**	**0.928**

Table 3
Results of Comparison among Different Models on *Weibo*.

Method	Acc.	Prec.	Recall	F_1
EANN	0.827	0.847	0.812	0.829
MVAE	0.824	**0.854**	0.769	0.809
SAFE	0.816	0.816	0.818	0.817
MFN	0.808	0.806	0.806	0.807
CALM	0.846	0.843	0.864	0.853
CAFE	0.840	0.825	0.851	0.837
SAAN	**0.853**	0.837	**0.872**	**0.854**

4.5. Ablation Study

We conduct an ablation study on the image (w/o visual) and text (w/o textual) from our multimodal model. In addition, we compare the performance of four variants with SAAN to further explore the importance of different modules. We ablate the self-attention-based module (w/o self-att), co-attention-based module (w/o co-att), and CLIP-guided similarity evaluator (w/o CLIP) by excising corresponding components from SAAN. w/o similarity loss is a variant keeping the CLIP-extracted features and fine-tuning process but dropping the similarity-based loss away to the final prediction. All the results are shown in Table 4.

We observe that the performance drops by 31.5% in accuracy and 40.8% in F1 score on *Twitter*, while only 1.6% in accuracy and 1.9% in F1 score on *Weibo*. In contrast, the decline of accuracy and F1 score is much more pronounced on Weibo when we ablate text. We consider that the reason might be the variability in the quality of different modality features in distinct datasets. For *Twitter*, characteristics in vision such as tampering traces are more significant than that in text. Furthermore, some semantic features such as writing style and syntax benefit more to *Weibo*.

The results of different variants indicate that (1) complete SAAN that integrates all components overperforms among all variants; (2) self-attention mechanism contributes most to the performance for Twitter; (3) for Weibo, CLIP-guided similarity evaluator is the most important component among others; (4) evaluating the mismatches of text and image can be beneficial to detecting fake news since adding similarity-based loss improves the accuracy and F1 score on both datasets.

Table 4
Ablation Study on Different Variants of SAAN.

Method	Twitter		Weibo	
	Acc.	F_1	Acc.	F_1
SAAN	**0.925**	**0.928**	**0.853**	**0.854**
-w/o Visual	0.610	0.507	0.837	0.835
-w/o Textual	0.905	0.905	0.615	0.615
-w/o self-att	0.872	0.860	0.845	0.850
-w/o co-att	0.906	0.903	0.850	0.851
-w/o CLIP	0.911	0.917	0.842	0.845
-w/o similarity loss	0.918	0.920	0.850	0.851

5. Conclusion

In this paper, we propose a multimodal method for fake news detection, named SAAN. It provides an available approach to integrating both the complementary and inconsistent information of news posts with text and images. We design an attention-based multimodal feature extractor to capture the correlation between modalities together with a CLIP-guided similarity evaluator to measure the inconsistency between the text and image. Experimental results show that SAAN can defeat all the multimodal baselines on two datasets.

6. References

[1] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 109:1–109:40, 2020.

[2] Martin Potthast and Johannes Kiesel and Kevin Reinartz and Janek Bevendorff and Benno Stein, "A stylometric inquiry into hyperparti- san and fake news," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 2018, pp. 231–240.

[3] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact- checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[4] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[5] Y. Chen, J. Sui, L. Hu, and W. Gong, "Attention-residual network with cnn for rumor detection," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1121– 1130.

[6] V. Vaibhav, R. M. Annasamy, and E. Hovy, "Do sentence interactions matter? leveraging sentence level representations for fake news classi- fication," in *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs@EMNLP 2019, Hong Kong, November 4, 2019*, 2019, pp. 134–139.

[7] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," *CoRR*, vol. abs/2003.05096, 2020.

[8] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*, 2019, pp. 39–47.

[9] J. Chen, Z. Wu, Z. Yang, H. Xie, F. L. Wang, and W. Liu, "Multimodal fusion network with latent topic memory for rumor detection," in *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*, 2021, pp. 1–6.

[10] Z. Wu, J. Chen, Z. Yang, H. Xie, F. L. Wang, and W. Liu, "Cross-modal attention network with orthogonal latent memory for rumor detection," in *Web Information Systems Engineering - WISE*

2021 - 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26-29, 2021, Proceedings, Part I. Springer, 2021, pp. 527–541.

[11] X. Zhou, J. Wu, and R. Zafarani, "SAFE: similarity-aware multi-modal fake news detection," in Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II, 2020, pp. 354–367.

[12] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, 2017, pp. 231–240.

[13] B. Han, X. Han, H. Zhang, J. Li, and X. Cao, "Fighting fake news: Two stream network for deepfake detection via learnable SRM," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 3, pp. 320–331, 2021.

[14] 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 Trans. Multim., vol. 24, pp. 3455–3468, 2022.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[17] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2015," 2015.

[18] Z. Jin, J. Cao, H. Guo, and Y. Z. andf Jiebo Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017, 2017, pp. 795–816.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimiza- tion," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[20] 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, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 849–857.

[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2019, pp. 2915–2921.

[22] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross- modal ambiguity learning for multimodal fake news detection," in *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 2022, pp. 2897–2905.