TT-BLIP: Enhancing Fake News Detection Using BLIP and Tri-Transformer

Eunjee Choi
School of Electrical Engineering
Korea University
Seoul, South Korea
eun09ji@korea.ac.kr

Jong-Kook Kim
School of Electrical Engineering
Korea University
Seoul, South Korea
jongkook@korea.ac.kr

***Abstract*—Detecting fake news has received a lot of attention. Many previous methods concatenate independently encoded unimodal data, ignoring the benefits of integrated multimodal information. Also, the absence of specialized feature extraction for text and images further limits these methods. This paper introduces an end-to-end model called TT-BLIP that applies the bootstrapping language-image pretraining for unified vision-language understanding and generation (BLIP) for three types of information: BERT and BLIP_{Txt} for text, ResNet and BLIP_{Img} for images, and bidirectional BLIP encoders for multimodal information. The Multimodal Tri-Transformer fuses tri-modal features using three types of multi-head attention mechanisms, ensuring integrated modalities for enhanced representations and improved multimodal data analysis. The experiments are performed using two fake news datasets, Weibo and Gossipcop. The results indicate TT-BLIP outperforms the state-of-the-art models.**

***Index Terms*—multimodal fusion, vision-language pretraining, fake news detection**

I. Introduction

The expansion of digital platforms has altered news consumption by broadening access and increasing exposure to misinformation, emphasizing the urgent need for enhanced fake news detection. Social media advancements have evolved news delivery from traditional text articles to multimedia narratives [1], integrating images and videos. Images provide visual information [2] that enhances the textual content and can attract more attention.

Fig. 1 shows two examples from the Weibo dataset. The first image and its text are easily verified as true. However, the second image and text, if viewed separately, may provide insufficient information to assess the authenticity of the news content. By examining both the image and text together the fabrication becomes clear, illustrating how combining image and textual information improves fake news detection.

The evolution of fake news detection techniques has transitioned from traditional approaches to more advanced deep learning methodologies. Previously, [3], [4] primarily utilized text content to discern between fake and real news. Building on this, Ma et al. [5] delved into the potential of deep neural networks for representing tweets, emphasizing temporal-linguistic attributes. Chen et al. [6] then augmented this approach by incorporating attention mechanisms into RNN

(a) Iris Grace, a British autistic girl with amazing painting talent.

(b) Eight party members and leading cadres were notified for disciplinary issues.

Fig. 1: Real news (a) and fake news (b) examples from the Weibo dataset

structures. Deep learning-based fake news detection highlights notable enhancements in performance over their conventional counterparts, attributing this to the superior feature extraction capabilities of the newer methods. Jin et al. [2] adopted an end-to-end approach by merging image, textual, and social context features, using attention mechanisms to inform their predictions. Meanwhile, Wang et al. [7] introduced an event discriminator that aimed to identify applicable features across various events and thereby enhancing the model adaptability to novel events. Existing models have a clear deficiency in specialized feature extraction for both text and images and using cross modal attention has shown its own set of challenges. These two main challenges made the effective detection of fake news more difficult.

This paper introduces Tri-Transformer BLIP (TT-BLIP) shown in Fig. 2, which is based on bootstrapping language-image pretraining for unified vision-language understanding and generation [8] to address the challenges above. Inspired by the use of three separate encoders in CLIP-MSA to capture distinct modality characteristics and enhance representations through inter-modal dynamics [9], our model employs BLIP [8], BERT [10], and ResNet [11] for feature extraction. It is a three-pathway model employing BERT and BLIP_{Txt} for text feature extraction, ResNet and BLIP_{Img} for image feature extraction and a pair of bidirectional BLIP encoders for image-text correlation information. The Multimodal Tri-Transformer fuses features from text, image, and image-

text modalities for fake news detection. The text modality employs self multi-head attention for internal analysis, while image and image-text modalities use cross-attention with text queries, aligning visual content with textual context. The process prioritizes text-driven analysis essential for evaluating multimodal fake news data. After the attention, each modality's output undergoes Multi-Layer Perceptron (MLP) transformation, ensuring data uniformity. The processed outputs from different modalities are then concatenated, leading to a unified and comprehensive representation. Experiments are conducted using two multimodal fake news datasets to assess the performance of TT-BLIP: Weibo [12] and Gossipcop [13]. The results demonstrate that TT-BLIP is the state-of-the-art model for fake news detection.

In summary, the main contributions of this paper are:

- TT-BLIP is proposed, which uses the pre-trained BLIP model for feature extraction to detect fake news.
- A novel fusion mechanism is introduced, which is a Multimodal Tri-Transformer. By fusing text, images, image-text features, the transformer captures the semantic information of all three modalities.
- The TT-BLIP is tested using two multimodal fake news, Weibo and Gossipcop, and results show that TT-BLIP performs better than the state-of-the-art multimodal detection models.

II. RELATED WORK

In the domain of Multimodal fake news detection, various strategies have emerged that emphasize extracting features from both images and texts from news articles. EANN [7] is designed to detect fake news from social media, particularly those involving event-invariant features. MVAE [14] is an end-to-end network developed for fake news detection that incorporates a bimodal variational autoencoder and a binary classifier. Spotfake by Singhal et al. [15] uses BERT [10] for textual feature extraction and VGG19 [16] for image features to efficiently detect fake news. They later refined their approach in Spotfake+ [17] to better identify fake news in full articles. SAFE [18] explored the relationship between textual and image information in news articles for fake news detection. CAFE [19] detects fake news from social platforms by adapting to cross-modal ambiguities and analyzing uni-modal and cross-modal features. LIIMR [20] selectively diminishes information from less significant modalities, while emphasizing and extracting related data from the dominant modality for each sample. MCAN [21] utilizes co-attention layers within its three sub-networks to merge textual and image features. This approach emphasizes capturing inter-dependencies across multimodal inputs for fake news detection.

Some methods employ more information from datasets. DistilBert by Allein et al. [22] utilize the latent representations of user-generated and shared content to detect disinformation in online news articles. By correlating user preferences and sharing behaviors, the model can efficiently differentiate between genuine and fake news without relying on user profiling during prediction. BDANN [23] tackles multimodal fake news detection on microblogging platforms by combining features from two modalities and eliminating event specific biases. FND-CLIP [24] applied the Contrastive Language-Image Pre-Training (CLIP, [25]) vision-language pretraining model to measure the correlation between images and texts and utilize different modalities for decision-making. FND-CLIP used two unimodal encoders, two pair-wise CLIP encoders, guides the network learning for various modalities, and adeptly aggregates text, image, and fused features with a modality-wise attention module.

Unlike previous models, TT-BLIP uses a three pathway model extracting features from image, text, and image-text multimodal information and the TT-BLIP model incorporates a Multimodal Tri-Transformer. This novel approach enables the fusion of text, images, and image-text features, offering a more comprehensive and effective method for utilizing multimodal information.

III. METHOD

A. Overview

In this paper, a new multimodal fake news detection model is proposed and Fig. 2 shows the overall architecture of TT-BLIP. TT-BLIP is composed of three modules: feature extraction module, feature fusion module, and fake news detector module. In the feature extraction module, image, text, and image-text features are extracted from text articles and associated images. In the fusion module, the MultiModal Tri-Transformer method is used to aggregate the output of the feature extraction module. Finally, in the fake news detector, the integrated features processed through the MultiModal Tri-Transformer are used to predict whether the content is real or fake.

B. Feature Extraction Layer

The feature extraction layer is comprised of text feature extraction layers, image feature extraction layers, and image-text feature extraction layers. The data used are denoted as x_{Img} for images and x_{Txt} for text.

1) ***Textual feature extractor***

: BERT [10] and BLIP_{Txt} [8] are used in parallel to capture a comprehensive representation of the textual data.

- **BERT-based Feature Extraction:** the pretrained BERT is used, bidirectional transformer model used for feature extraction. From this, the feature representation $z_{\text{t1}} = f_{\text{bert}}(x_{\text{Txt}})$ is obtained. The strength of BERT is its bidirectional training that enables it to understand words in their context and thus enhancing textual feature extraction.
- **BLIP_{Txt}-based Feature Extraction:** BLIP_{Txt} is used to derive $z_{\text{t2}} = f_{\text{blip\_T}}(x_{\text{Img}}, x_{\text{Txt}})$. BLIP is inherently designed to accept both text and images as inputs simultaneously. However, to ensure that only textual information is considered, an image tensor (dimension: $224 \times 224 \times 3$) filled with zeros is inserted for the image data. This ensures that BLIP focuses solely on the textual data and extracts relevant features accordingly.

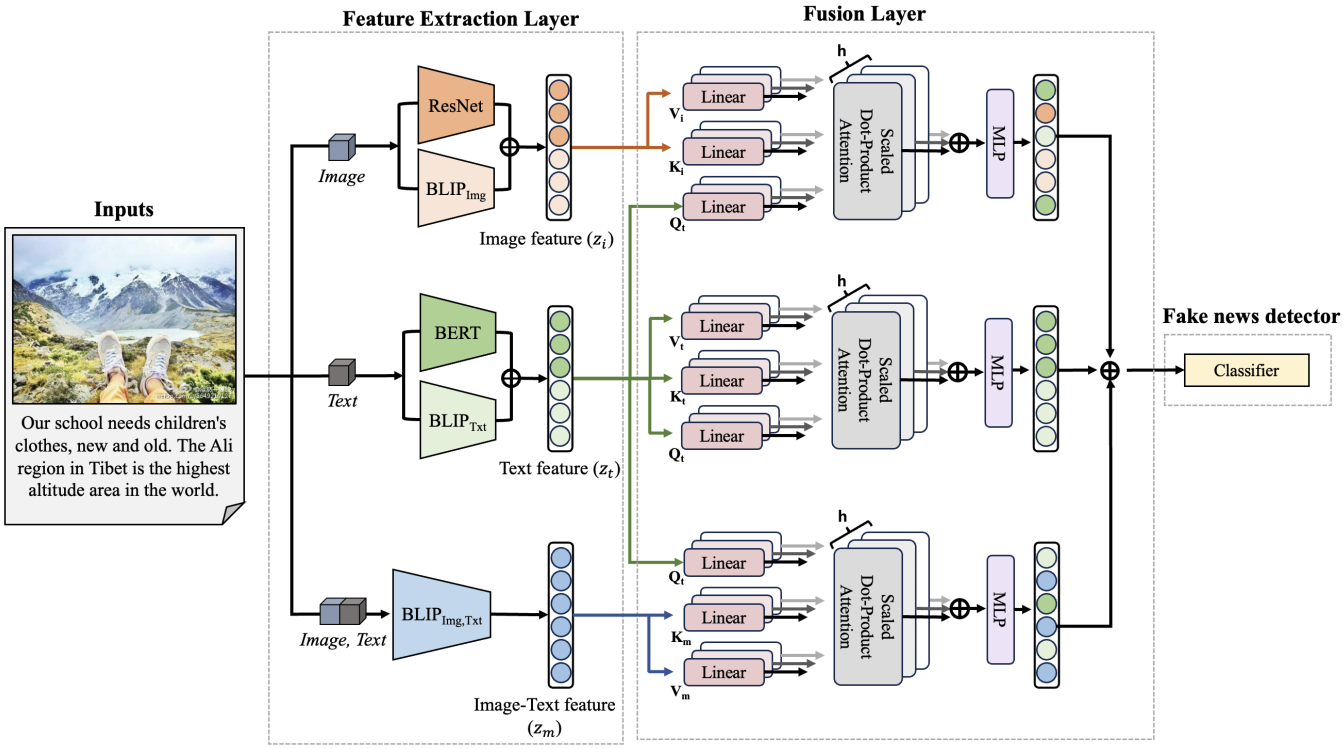


Fig. 2: The architecture of the proposed TT-BLIP.

*2) **Image feature extractor***

For the image data, ResNet [11] and BLIP$_{\text{Img}}$ are used in parallel.

- **ResNet-based Feature Extraction:** the pretrained ResNet model is employed for image feature extraction. The feature representation $z_{\text{i1}} = f_{\text{resnet}}(x_{\text{Img}})$ is obtained. This choice ensures that our model can effectively capture intricate patterns and features from images.
- **BLIP$_{\text{Img}}$-based Feature Extraction:** the BLIP$_{\text{Img}}$ is used to derive $z_{\text{i2}} = f_{\text{blip\_i}}(x_{\text{Img}}, x_{\text{Txt}})$, which serves as an alternative approach for image feature extraction. BLIP is designed to handle both text and images as input simultaneously. However, to focus solely on extracting image features in this work a dummy text is input. This process ensures that the textual aspect does not influence the extraction of image features.

The two image and two textual feature representations are then integrated into single image and textual representations, respectively. For image features, this integrated representation is denoted as z_{i}, and for text features, it is denoted as z_{t}. These integrated features are achieved by concatenating the two respective feature information shown as $z_{\text{i}} = [z_{i1}; z_{i2}]$ for the image features and $z_{\text{t}} = [z_{t1}; z_{t2}]$ for text features. These are then used in subsequent stages of the proposed model.

*3) **Image-text feature extractor***

: the pretrained BLIP model is used to extract image-text features, aiming to combine image and textual information. This process is essential to create a comprehensive representation that captures the interaction between image and text data. The resulting Image-text feature representation integrates information from both modalities and can be expressed as follows: $z_{\text{m}} = f_{\text{blip}}(x_{\text{Img}}, x_{\text{Txt}})$. Here, z_{m} represents the integrated feature representation.

C. Feature Fusion Layer

In the feature fusion layer (Fig 3), the MultiModal Tri-Transformer is introduced and this idea comes from several MultiModal Transformers [26]–[28]. The Tri-Transformer uses three types of multi-attention [29] mechanisms. It applies cross-modal attention between text and both image-text and image modalities. This ensures the text modality, which is crucial for the task, is emphasized more than the others while keeping the image and image-text channels independent. For text alone, it employs self multi-head attention to enhance textual analysis. All components within this fusion process, including the attention mechanisms and MLP (Multi Layer Perceptron) layer, are trained from scratch to ensure integrated and uniform data processing.

$$\text{C\_Attention}_{i-t}(Q, K, V) = \text{softmax}\left(\frac{Q_t K_i^T}{\sqrt{d_h}}\right) V_i \quad (1)$$

$$\text{C\_Attention}_{m-t}(Q, K, V) = \text{softmax}\left(\frac{Q_t K_m^T}{\sqrt{d_h}}\right) V_m \quad (2)$$

$$\text{S\_Attention}_t(Q, K, V) = \text{softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_h}}\right) V_t \quad (3)$$

Let Q_t represent the query vector for text modality, used to extract relevant information. The key vectors for the text, image, and image-text modalities are denoted by K_t, K_i, and K_m, determining data relevance. The value vectors for these modalities are represented by V_t, V_i, and V_m, holding the actual data to be retrieved. The softmax$(\cdot)$ function refers to the softmax function. The term d_h indicates the dimensionality of each modality, and T stands for transpose.

In the Transformer architecture, several attentions are computed in parallel, with each attention's output known as a head. The num^{th} head is computed as:

$$\text{head}_*^{num} = \text{Attention}_*\left(Q_t W_{num}^{Q_t}, K_* W_{num}^{K_*}, V_* W_{num}^{V_*}\right) \quad (4)$$

where $\mathbf{W}_{num}^{Qt} \in \mathbb{R}^{d_t \times d_q}$ is the weight matrix of Q_t when computing the head of the num^{th} text modality; $\mathbf{W}_{num}^{K*} \in \mathbb{R}^{d_* \times d_k}$ be the weight matrix of K_* when computing the head of the num^{th} $*$ modality; and $\mathbf{W}_{num}^{V*} \in \mathbb{R}^{d_* \times d_v}$ be the weight matrix of V_* when computing the head of the num^{th} $*$ modality, where $* \in \{i, t, m\}$.

Following this, all heads of the $*$ modalities are connected, denoted as Y_*, as follows:

$$\begin{aligned} Y_* &= \text{MultiHead}(Q_t, K_*, V_*) \\ &= \text{Concat}(\text{head}_*^1, \text{head}_*^2, \ldots, \text{head}_*^n)\mathbf{W}_{O_*}, \end{aligned} \quad (5)$$

Where $\mathbf{W}O_*$ is the weight matrix multiplied after splicing the head of $*$ modalities and n denotes the number of self-attention heads used. Thus, the text-based image representation f_{it}, the text-based image-text representation f_{mt}, and text representation f_t can be obtained as follows:

$$f_{at} = \text{MultiHead}(Y_i; \theta_{c-\text{att}}^i), \quad (6)$$

$$f_{vt} = \text{MultiHead}(Y_m, \theta_{c-\text{att}}^m), \quad (7)$$

$$f_t = \text{MultiHead}(Y_t, \theta_{\text{att}}^t) \quad (8)$$

where $\theta_{\text{att}}^i = \{\mathbf{W}_Q^i, \mathbf{W}_K^i, \mathbf{W}_V^i, \mathbf{W}_O^i\}$ and $\theta_{\text{att}}^m = \{\mathbf{W}_Q^m, \mathbf{W}_K^m, \mathbf{W}_V^m, \mathbf{W}_O^m\}$ represent the main hyperparameters required for the cross-attention module. $\theta_{\text{att}}^t = \{\mathbf{W}_Q^t, \mathbf{W}_K^t, \mathbf{W}_V^t, \mathbf{W}_O^t\}$ is used for the self-attention module.

After the multi-head attention step, each modality's output is passed through its own MLP (Multi Layer Perceptron) layer. The MLP layer perform linear transformations on data, integrating features extracted from previous layers to ensure a uniform format and facilitate the generation of the final output. The processed outputs are finally concatenated into a unified tensor.

D. Fake News Detector

A fake news detector is used to verify the authenticity of multimodal news articles. The combined data from the fusion layer is sent to a classifier, which comprises three linear layers with ReLU activations, batch normalization, and outputs binary classification for fake news detection. The label set is denoted by $\mathcal{Y}$. Specifically, a news article that is classified as 'fake' is assigned a label of 1, and 0 otherwise. To assess the difference between the predicted and actual labels, the cross-entropy loss is used and is formulated as:

$$L_{\text{cls}} = -\sum_{(p, y_i) \in (\mathcal{P}, \mathcal{Y})} [y \cdot \log(\hat{y}_i) + (1 - y) \cdot \log(1 - \hat{y}_i)] \quad (9)$$

IV. EXPERIMENTS AND RESULTS

A. Dataset

1) Weibo

The Weibo dataset was introduced by Jin et al. [12]. The dataset aggregates authentic news from esteemed Chinese news outlets, with the Xinhua News Agency being a prime example. The deceptive news articles were crawled and authenticated by Weibo's official rumor verification system from May 2012 to January 2016. This system encourages general users to report potential misinformation, which is then assessed by a panel of reputable users. The experimental training set consists of 6,137 news articles, where there are 2,802 fake and 3,335 real news pieces, while the testing set comprises of 833 fake and 852 real news.

2) Gossipcop

Gossipcop dataset is used in this paper ([13]). There are 10010 news articles in the training set, where 2036 are fake and 7974 are real news pieces. The testing set consists of 545 fake and 2285 real news.

B. Experimental Settings

For text feature extraction, the pretrained BERT model which is based on Chinese was used for the Weibo dataset, and the "bert-base-uncased" model for the Gossipcop dataset. In terms of image feature extraction, the input image size is set to 224 × 224 and ResNet [11] was used. Lastly, image-text features are extracted by pairing images and text and processing them with the pretrained BLIP [8] model. Because the BLIP is pretrained using English text, for the Weibo dataset the Google Translation API is utilized to translate Chinese texts into English. The model is trained using the Adam optimizer with default parameters, a learning rate of 1×10^{-3} and a batch size of 64 for 50 epochs.

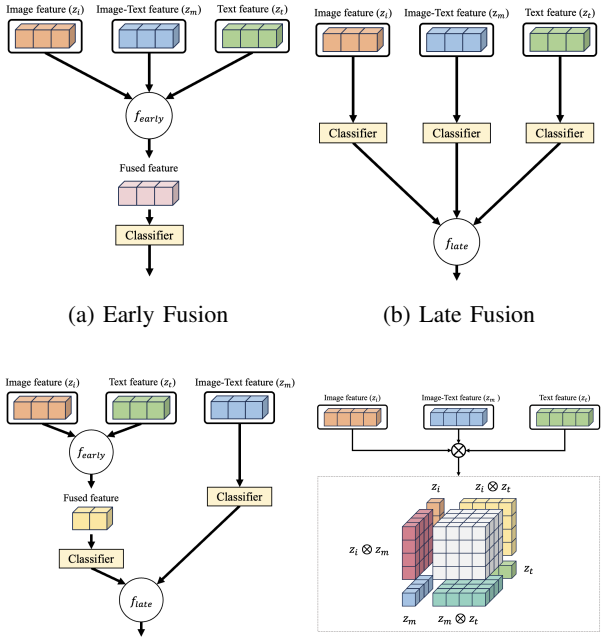
C. Results and Analysis

TT-BLIP is compared to the state-of-the-art models which are presented in Table I to validate the performance. The experiments are performed using the Weibo and the Gossipcop datasets. For the evaluation, accuracy, precision, recall, and F1 scores are calculated for both real and fake news. Several strategies, such as EANN and Spotfake, use concatenation or attention mechanisms for feature fusion but often face correlation deficiencies because features remain in distinct semantic spaces. CAFE attempts to mitigate this through cross-modal alignment, aligning texts and images within a unified

TABLE I: Experimental Results on Weibo and Gossipcop Datasets
'-' symbol indicates that the results are not available from the original paper.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
Weibo	EANN [7]	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	MVAE [14]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	Spotfake [15]	0.892	0.902	**0.964**	0.932	0.847	0.656	0.739
	SAFE [18]	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	BDANN [23]	0.821	0.790	0.610	0.690	0.830	0.920	0.870
	LIIMR [20]	0.900	0.882	0.823	0.847	0.908	0.941	0.925
	MCAN [21]	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	CAFE [19]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP [24]	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	TT-BLIP(VGG)	0.960	0.963	0.952	0.957	**0.951**	0.963	0.957
	TT-BLIP(XLNet)	0.957	0.963	0.952	0.957	**0.951**	0.963	0.957
	TT-BLIP(ours)	**0.961**	**0.979**	0.944	**0.961**	0.944	**0.980**	**0.962**
Gossipcop	Spotfake+ [17]	0.856	-	-	-	-	-	-
	LSTM-ATT [30]	0.842	-	-	-	0.839	0.842	0.821
	SAFE [18]	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	DistilBert [22]	0.857	0.805	0.527	0.637	0.866	**0.960**	0.911
	CAFE [19]	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	FND-CLIP [24]	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	TT-BLIP(VGG)	0.846	**0.842**	0.552	0.667	0.847	**0.960**	0.900
	TT-BLIP(XLNet)	0.865	0.655	**0.792**	**0.717**	0.875	0.933	0.903
	TT-BLIP(ours)	**0.885**	0.737	0.596	0.659	**0.910**	0.950	**0.930**

semantic space. Yet, its effectiveness is limited by small datasets and general labels, resulting in persistent semantic gaps between textual and image features. FND-CLIP uses two pre-trained CLIP encoders to extract the representations from the image and text. TT-BLIP achieved the highest accuracy of 96.1% and 88.5% in detecting fake news mainly due to the following reasons: 1) the TT-BLIP architecture combines ResNet and BLIP_{Img} for image data and BERT and BLIP_{Txt} for text processing, while using bidirectional BLIP encoders for correlation information. ResNet is used for extracting complex features from images, aiding in news image analysis. BERT is used for understanding language nuances, important for analyzing news text. Experiments with alternatives like TT-BLIP(VGG) (employing VGG [16] instead of ResNet) and TT-BLIP(XLNet) (using XLNet [31] instead of BERT) demonstrated the superiority of the original TT-BLIP setup. 2) The MultiModal Tri-Transformer uses attention mechanisms across multiple modalities, ensuring a comprehensive and integrated representation of the data. This approach allows for a more comprehensive understanding of context and improves feature extraction through multi-head attention analyzing data from various perspectives. By focusing on important text and merging features from all modalities, this architecture improves data representation, essential for better classification accuracy.



(a) Early Fusion (b) Late Fusion

(c) Hybrid Fusion (d) Tensor Fusion

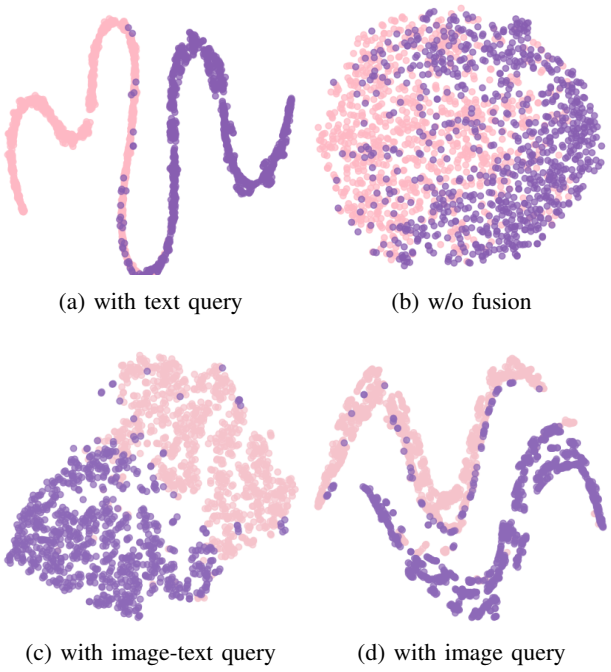
Fig. 3: Architecture of different fusion strategies for Multimodal fake news detection

D. Comparison of Fusion Methods

In Table II, the TT-BLIP model, utilizing a multi-head attention mechanism, is compared against traditional fusion approaches such as early fusion [32], late fusion [32], hybrid fusion [33], and tensor fusion [34] on the Weibo and Gossipcop datasets. Early Fusion initially combines image, text, and image-text features, then applies classification to this unified dataset. Late Fusion aggregates results from different classifiers, each trained on separate modalities like image features, text features, and image-text features. Hybrid Fusion merges

TABLE II: Comparative Analysis of Fusion Techniques for Fake News Detection on Weibo and Gossipcop Datasets

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
Weibo	Early fusion [32]	0.792	0.828	0.742	0.783	0.761	0.843	0.800
	Late fusion [32]	0.806	0.840	0.761	0.799	0.777	0.852	0.813
	Hybrid fusion [33]	0.815	0.850	0.770	0.808	0.785	0.861	0.821
	Tensor fusion [34]	0.821	0.876	0.786	0.829	0.766	0.864	0.812
	TT-BLIP	**0.961**	**0.979**	**0.944**	**0.961**	**0.944**	**0.980**	**0.962**
Gossipcop	Early fusion [32]	0.634	0.277	0.560	0.370	0.861	0.651	0.742
	Late fusion [32]	0.713	0.244	0.233	0.238	0.819	0.828	0.823
	Hybrid fusion [33]	0.853	0.613	**0.642**	0.627	**0.914**	0.903	0.909
	Tensor fusion [34]	0.868	**0.754**	0.466	0.576	0.883	**0.964**	0.922
	TT-BLIP	**0.885**	0.737	0.596	**0.659**	0.910	0.950	**0.930**



(a) with text query (b) w/o fusion

(c) with image-text query (d) with image query

Fig. 4: t-SNE visualization of extracted features from the Weibo test set using TT-BLIP. Each color represents a distinct label grouping.

image, text, and image-text features by employing both early and late fusion methods. Tensor Fusion combines image, text, and image-text features using the outer product, enhancing interaction between modalities for improved classification. The architectures of traditional fusion methods are depicted in Fig. 3, demonstrating structural differences and highlighting the innovation of the TT-BLIP model.

In Table II, TT-BLIP outperforms traditional methods significantly. For the Weibo dataset, TT-BLIP significantly outperformed the traditional methods on Weibo with an accuracy of 96.1% and achieved higher precision (97.9%) and F1 scores (96.1% and 96.2%) for fake news detection. Other methods did not perform as well, with the highest accuracy among them being only 82.1%. On the Gossipcop dataset, TT-BLIP was superior with 88.5% accuracy, better than tensor fusion at 86.8%. In terms of precision and recall for fake news detection, TT-BLIP performs competitively, though not leading, with figures at 73.7% and 59.6% respectively. This indicates that while TT-BLIP excels in overall accuracy and F1 score, it encounters competition from methods like Hybrid and Tensor Fusion in specific metrics.

In Fig. 4, t-SNE visualizations analyze features processed by TT-BLIP in different settings within the MultiModal Tri-Transformer's feature fusion mechanism, using the test set of Weibo. The visualizations compare before and after fusion, utilizing different types of queries: text, image-text, and image. The dots with the same color depicts data with same label. In Fig. 4, the visualizations show that before fusion (Fig. 4 (b)) data points are indicating a lack of clear separation between fake and real news instances. Applying TT-BLIP, there's a improvement in clustering, the real and fake news instances are now grouped more distinctly. This enhanced separation facilitates a more effective identification of fake news by allowing the model to better integrate the underlying patterns within the data. In Fig. 4, which displays various query configurations, a comparative analysis of (a), (c), and (d) shows differences in data clustering effectiveness. Specifically, configurations employing image-text and image queries, as shown in Fig. 4 (c) and (d), result in less defined separations between clusters. This outcome indicates that methods relying solely on image data are less effective for classification. The reduced clarity in clustering with image or image-text queries emphasizes the significant role of text in creating a clear and distinct feature space. Textual content offers crucial context, improving accuracy in differentiating categories, particularly in tasks like fake news detection. This comparison shows that incorporating textual information improves the model's classification efficacy.

The distinct clustering with text queries demonstrates the value of text in improving the model's capacity to distinguish between categories, essential for tasks such as fake news detection. Fig. 4 confirms the benefit of prioritizing text in multimodal feature fusion, yielding more distinct feature representations and supporting better classification outcomes.

TABLE III: Ablation experimental results of TT-BLIP

Dataset	ResNet	BLIP_{Img}	BERT	BLIP_{Txt}	$\text{BLIP}_{\text{Img,Txt}}$	Fusion	Accuracy	F1 Score	
								Fake News	Real News
Weibo	✓	✓					0.659	0.684	0.629
			✓	✓			0.837	0.836	0.839
		✓		✓	✓	✓	0.858	0.846	0.869
	✓		✓			✓	0.912	0.911	0.912
					✓		0.747	0.747	0.747
	✓	✓	✓	✓	✓		0.852	0.853	0.851
		✓	✓	✓	✓	✓	0.959	0.959	0.959
	✓	✓		✓	✓	✓	0.936	0.937	0.935
	✓	✓	✓	✓	✓	✓	**0.961**	**0.961**	**0.962**
Gossipcop	✓	✓					0.798	0.861	0.632
			✓	✓			0.846	0.892	0.733
		✓		✓	✓	✓	0.846	**0.895**	0.714
	✓		✓			✓	0.848	0.715	0.902
					✓		0.837	0.746	0.879
	✓	✓	✓	✓	✓		0.837	0.886	0.712
		✓	✓	✓	✓	✓	0.856	0.717	0.903
	✓	✓		✓	✓	✓	0.846	0.692	0.897
	✓	✓	✓	✓	✓	✓	**0.885**	0.659	**0.930**

E. Ablation study

Some variants and components of the model were tested to identify the importance of the proposed model. The results are shown in Table III. If the fusion module was not used, the features were directly concatenated. The effectiveness of each components is measured through two metrics: Accuracy and F1 Score, with separate F1 Scores for Fake News and Real News. The Text-only model, utilizing BERT and BLIP_{Txt} for textual analysis, proved more effective than the Image-only model which employed ResNet and BLIP_{Img} for image-based feature extraction. This indicates that textual features are more crucial in the classification of fake news. Integrating BLIP_{Img}, BLIP_{Txt}, and $\text{BLIP}_{\text{Img,Txt}}$ slightly outperformed the Text-only model, indicating the benefits of incorporating image features and textual data. The performance of the $\text{BLIP}_{\text{Img,Txt}}$ only model, which uses only $\text{BLIP}_{\text{Img,Txt}}$ combined feature without any fusion, did not reach the accuracy levels of the integrating BLIP_{Img}, BLIP_{Txt}, and $\text{BLIP}_{\text{Img,Txt}}$ method. This highlights the necessity for more integrated usage of features. The removal of the ResNet module resulted in a minor decline in performance, implying its lesser importance compared to other components. In contrast, the exclusion of the BERT component led to a significant reduction in effectiveness, emphasizing the critical role BERT plays in text processing. Models lacking fusion methods were less effective than TT-BLIP, underlining the importance of advanced fusion techniques. The complete TT-BLIP model, integrating ResNet, BLIP_{Img}, BERT, BLIP_{Txt}, $\text{BLIP}_{\text{Img,Txt}}$, and fusion techniques, showed the highest accuracy in classifying fake news.

V. Conclusion

This study presented the Tri-Transformer BLIP (TT-BLIP), which uses the bootstrapping language-image pretraining (BLIP), that is used for enhanced vision-language understanding, and BERT for text feature extraction, BLIP and ResNet for image feature extraction, BLIP for image-text feature extraction and the Multimodal Tri-Transformer for the fusion of the feature from different modalities. A key contribution is the three-pathway structure of TT-BLIP, which processes text and images independently while also learning the correlation of the two different modals and the Multimodal Tri-Transformer that provides an efficient fusion mechanism to integrate information across these three modalities. The TT-BLIP outperformed the previous state of the art fake news detection models in accuracy by 5.4% for the Weibo dataset (90.7% versus 96.1%) and 0.5% for the Gossipcop (88% versus 88.5%).

VI. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1B04933156).

References

[1] Sakshini Hangloo and Bhavna Arora, "Combating multimodal fake news on social media: methods, datasets, and future perspective," *Multimedia systems*, vol. 28, no. 6, pp. 2391–2422, 2022.

[2] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[3] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[4] Nadia K Conroy, Victoria L Rubin, and Yimin Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the association for information science and technology*, vol. 52, no. 1, pp. 1–4, 2015.

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

[6] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber,*

PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22. Springer, 2018, pp. 40–52.

[7] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing 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*, 2018, pp. 849–857.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12888–12900.

[9] Qi Huang, Pingting Cai, Tanyue Nie, and Jinshan Zeng, "Clip-msa: Incorporating inter-modal dynamics and common knowledge to multimodal sentiment analysis with clip," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8145–8149.

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[13] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[14] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915–2921.

[15] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 2019, pp. 39–47.

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

[17] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, pp. 13915–13916.

[18] Xinyi Zhou, Jindi Wu, and Reza Zafarani, "Safe: similarity-aware multi-modal fake news detection (2020)," *Preprint. arXiv*, vol. 200304981, pp. 2, 2020.

[19] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2897–2905.

[20] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru, "Leveraging intra and inter modality relationship for multimodal fake news detection," in *Companion Proceedings of the Web Conference 2022*, 2022, pp. 726–734.

[21] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569.

[22] Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta, "Like article, like audience: Enforcing multimodal correlations for disinformation detection," *arXiv preprint arXiv:2108.13892*, 2021.

[23] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[24] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang, "Multimodal fake news detection via clip-guided learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2825–2830.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[26] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[27] Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang, "Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation," *IEEE Transactions on Multimedia*, 2023.

[28] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 904–915.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] Jun Lin, Glenna Tremblay-Taylor, Guanyi Mou, Di You, and Kyumin Lee, "Detecting fake news articles," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3021–3025.

[31] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[32] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.

[33] Zhiyong Wu, Lianhong Cai, and Helen Meng, "Multi-level fusion of audio and visual features for speaker identification," in *Advances in Biometrics: International Conference, ICB 2006, Hong Kong, China, January 5-7, 2006. Proceedings*. Springer, 2005, pp. 493–499.

[34] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.