Leveraging Multimodal Fusion for Advanced Fake News Detection

Anonymous ACL submission

Abstract

Detecting multimodal fake news is imperative for maintaining social media security and safeguarding community well-being. Existing detection approaches often fall short in adequately considering the nuanced context of social media and fail to fully utilize various modalities such as metadata, resulting in a significant gap. In this paper, we propose a novel and efficient model that integrates both textual global features and local features. This model captures semantic relationships within the text and utilizes a global corpus representation to align with the complex context of social media. We further enhance feature connectivity by employing a multilevel fusion technique that integrates visual and metadata information. Extensive experiments demonstrate that our method achieves state-of-the-art performance across all classification tasks using Fakeddit, the largest multimodal fake news dataset, underscoring its effectiveness.

1 Introduction

In our increasingly digitalized society, the proliferation of fake news and disinformation has extended across various channels, including journalism, news reporting, and social media. This surge in false information has led to numerous issues, such as inciting unfounded fears during health crises like the COVID-19 pandemic (Rocha et al., 2021). The impact of fake news is exacerbated by the widespread use of popular social media platforms and online sources that lack robust fact-checking measures or third-party filters, enabling individuals to disseminate false information on a large scale with relative ease. Moreover, social media platforms provide a versatile medium where users can share textual content alongside images, adding complexity to the detection of fake news (Nakamura et al., 2020). Therefore, research in fake news detection is crucial for ensuring the well-being of our community.

To effectively detect online fake news, many researchers have proposed using deep learning techniques for automatic detection. Early efforts primarily focused on monomodal detection, which concentrated on the textual aspect of online information, as fake news initially appeared predominantly in text form. However, the evolution of social media now allows users to incorporate images and videos to enhance their content, providing a medium for fake news that may rely heavily on visual elements, with text serving as a supplementary component. As a result, multimodal fake news detection methods have been developed to address this challenge. Additionally, several studies have highlighted the importance of comments and metadata, such as the number of comments and followers, as valuable sources of information for determining the authenticity of a piece of information. This introduces another crucial modality alongside text and images.

Despite these advancements, significant challenges remain in the field of multimodal fake news detection. Firstly, the feature extraction process for textual content, including the original post and associated comments, often lacks a nuanced understanding of the context. Secondly, few studies have fully leveraged all available modalities within fake news datasets; most focus solely on the original text and images, paying insufficient attention to associated comments and metadata. This oversight creates a noteworthy gap in the research that warrants further investigation.

In this paper, we propose an effective model to bridge these gaps, as illustrated in Fig. 1. Our model comprises three integral components, encompassing all valuable modalities within the dataset. In the "Textual Input" section, we meticulously analyze the content based on the original post and associated comments. We utilize BERT to extract **local embeddings** and employ a vocabulary graph to establish global relationships within

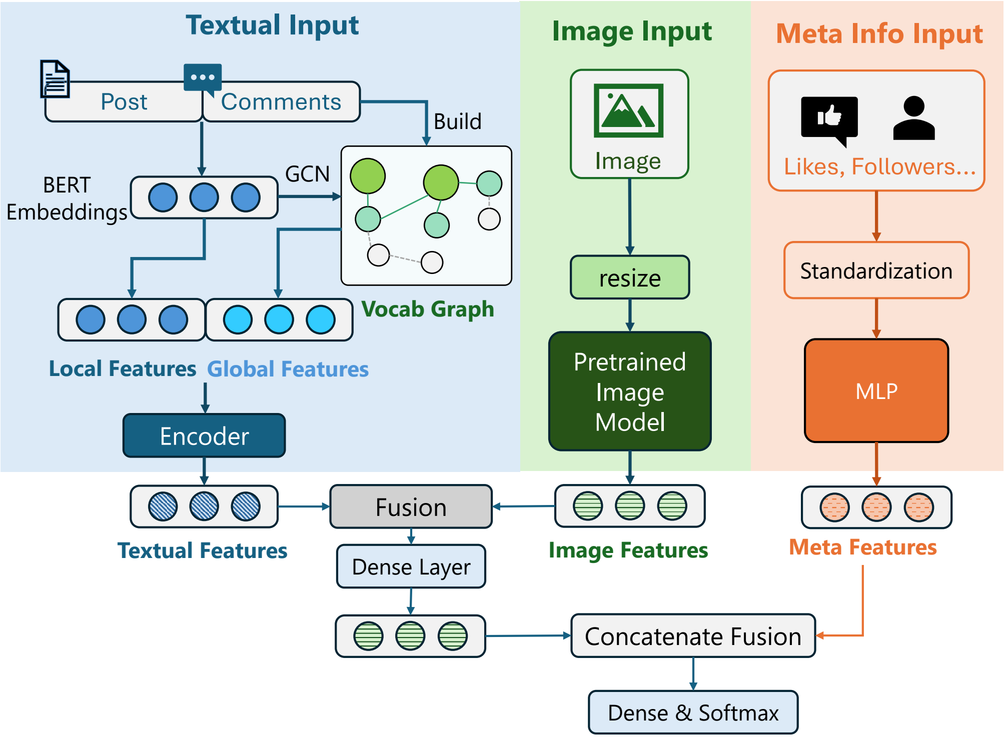

Figure 1: Overall Structure of Our Method

the dataset's lexicon, generating **global embeddings** for the textual content. By concatenating the local and global features, we feed them into a transformer encoder to fuse them using an attention mechanism, resulting in comprehensive textual features for the model. In the "Image Input" section, we employ a pretrained image model to extract visual features from the resized input image. Lastly, in the "Meta Info Input" section, we utilize the post's metadata information, such as the number of likes and followers. Since these numbers are quantifiable, we apply standardization and use a multilayer perceptron (MLP) to obtain meta features. Our multimodal feature fusion approach employs a multilevel fusion strategy. Initially, we fuse textual and image features through techniques such as concatenation, averaging, maximization, or addition. Subsequently, we apply concatenation fusion to incorporate meta features after compressing the feature dimensions. The experimental results on the widely-used Fakeddit dataset (Nakamura et al., 2020) demonstrate that our model achieves state-of-the-art (SOTA) performance across all classification tasks (2-way, 3-way, and 6-way), underscoring the effectiveness of our approach. Furthermore, the ablation tests provide additional validation for the significance of each module in our model.

Our contributions can be summarized as follows:

- We introduce a novel structure for multimodal fake news detection, incorporating a vocabulary graph and text global representations, which enriches contextual information.
- We utilize all available modalities within the dataset, employing a meticulously designed multilevel feature fusion structure to effectively leverage the information and enhance fake news identification.
- Our model achieves state-of-the-art performance across all classification tasks in the dataset, with a comprehensive discussion on how different fusion methods impact performance.

2 Related Works

Based on the modalities employed in fake news detection, the related literature can be categorized into two groups: monomodal approaches and multimodal approaches.

2.1 Monomodal Fake News Detection

Many early monomodal approaches concentrate on the textual content of online posts. Lin et al. (2019) proposed a novel rumor detection method based on a hierarchical recurrent convolutional neural network and a bidirectional GRU network with an attention mechanism, which integrates contextual features and captures time-period information. Ma et al. (2018) introduced a text-based method for rumor detection by combining propagation trees with recurrent neural networks. Comments are also considered a part of the text modality. Xu et al. (2022) concatenated the original post and associated comments to form a single long text, which is then segmented into shorter chunks more suitable for BERT-based vectorization and further classification. However, the drawback of monomodal approaches is that they may not adapt well to the evolving social media landscape, where new instances of fake information are increasingly conveyed through images, with text serving only as a supplementary component.

2.2 Multimodal Fake News Detection

Fake information can manifest across various modalities. Singhal et al. (2019) proposed a multimodal framework for fake news detection that employs a language transformer and visual models. Modality-specific feature representations are fused through concatenation and fed into a binary classifier. Similarly, Nakamura et al. (2020) utilized a two-stream network to process textual and visual information. A bidirectional BERT encodes text data, while a ResNet50 model processes visual data. The resulting embeddings are fused by taking their element-wise maximum. In contrast, Dong et al. (2018) utilized textual information and user-based metadata, such as age and the number of followers. These features are combined and processed by an attention-based bidirectional Gated Recurrent Unit network. The extracted features are then fed into a unified attention model, with patterns in attention distribution leveraged to detect fake news.

Only a few approaches leverage textual, visual, and metadata simultaneously. Cui et al. (2019) proposed a multistream architecture with an adversarial loss that individually influences all three network branches. The fused features are fed into a fully connected layer followed by a softmax to obtain likelihoods for fake news. Kirchknopf et al. (2021) proposed a multimodal network architecture enabling different levels and types of information fusion. They also utilize secondary information, such as user comments and metadata, along with text and images, and fuse this information while accounting for the specific intrinsic structures of the modalities.

Compared to these multimodal detection approaches, our method delves deeper into the detailed modalities of the original post and associated comments. We use a vocabulary graph and graph convolution to represent global features, which differ from the local features extracted in the related works mentioned above. Furthermore, our model leverages all available modalities with a multilevel fusion approach to further enhance detection accuracy.

3 Methodology

The overall structure of our method is presented in Fig. 1. Our approach can be divided into two main components: multimodal feature extraction and multimodal feature fusion.

3.1 Multimodal Feature Extraction

3.1.1 Integrating Global Features with Local Features

While BERT excels at learning local semantic features of text, it faces limitations in encoding information with remote dependencies due to constraints on input text length (Lu et al., 2020). This inability to capture word features across the entire corpus poses challenges in nuanced sentiment analysis, leading to potential misclassification. This highlights the importance of extracting global features alongside local features for effective fake news detection, especially in contexts where expressions are nuanced, as is common on social media.

For instance, consider the following comment extracted from a social media post:

Although it seems to be true, few have tried to find out the truth of the case.

In this comment, both affirmative and negative sentiments regarding whether the event is fake news are expressed. However, the affirmative sentiment that views the event as fake news is expressed indirectly, using phrases like "few ... find out the truth," while the negative sentiment is conveyed more directly, as in "seems to be true." In such contexts, if the language representation model does not associate this nuanced affirmative sentiment with the concept of "fake news," the classifier may underestimate this strong viewpoint, resulting in the comment being incorrectly classified as "not fake news."

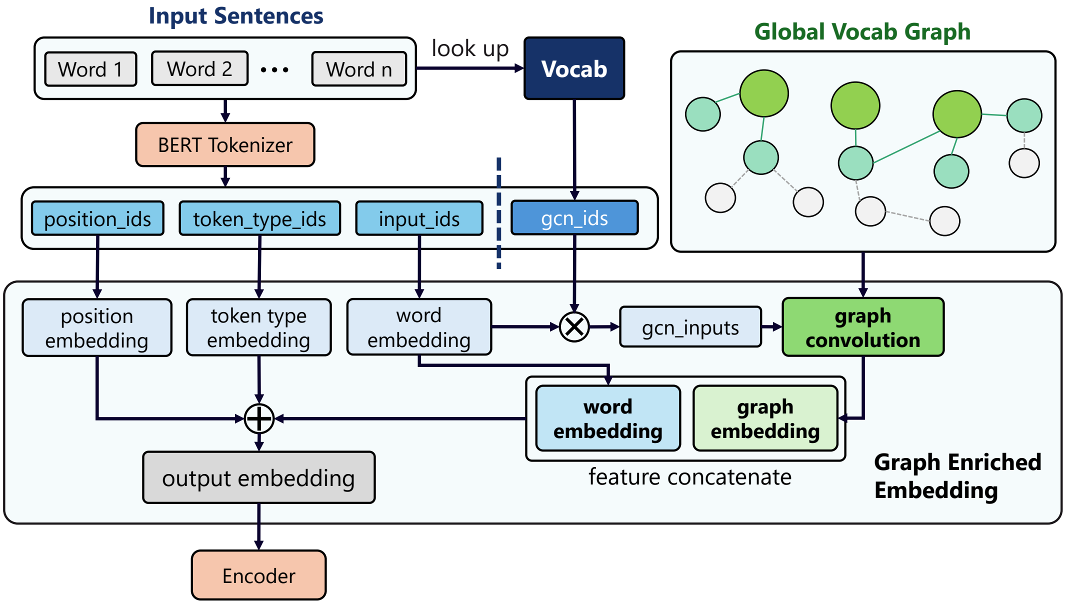

Figure 2: Local and Global Textual Feature Extraction

Recent research highlights the Graph Convolutional Network (GCN) and its variant, the Vocabulary Graph Convolutional Network (VGCN), for their effectiveness in extracting global features (Kipf and Welling, 2017; Lu et al., 2020). VGCN connects words in the language through a graph, enabling convolutional operations on neighboring nodes to integrate representations of words with those of their neighbors. This captures global relationships within specific domain language to some extent. However, traditional graph convolution, which focuses solely on global lexical information, overlooks word order in sequences, which is crucial for sentence understanding. Relying solely on global features is insufficient for accurately representing a sentence. Therefore, an appropriate structure is needed to fuse the two types of textual features extracted. Fig. 2 illustrates how we leverage both local and global features to represent textual features effectively.

We begin by concatenating the original post and comments to create an extended text input. Using the BERT tokenizer and embedding layer, we generate three types of embeddings: position, token type, and word embeddings. The word embeddings capture local features since they are learned solely within the input text range. These word embeddings serve as the input features for the VGCN model, where graph convolution is performed on the global vocabulary graph to derive the global graph embeddings. The detailed mathematical procedures are outlined below:

Assume a text segment is represented by a row vector x, where each element corresponds to a word node in the global vocabulary graph. The definition of a single-layer convolution operation on this text row vector is given by (Lu et al., 2020):

$$h = (Ax^T)^T W = xAW$$

Here, A represents the original neighbor matrix of the entire graph and is a symmetric matrix. In this expression, the xA operation extracts node information from the global vocabulary graph related to the input vector x, and W is a hidden state weight matrix. Through this operation, a global feature representation of the input row vector x is obtained.

Now, replacing the row vector x with the BERT word embeddings and considering a two-layer graph convolution structure, the obtained graph embeddings can be presented as:

$$G = ReLU(X_{bev} A_{vv} W_{vh}) W_{hg}$$

Here, b represents the batch size, e is the dimension of the BERT embedding, v is the size of the vocabulary, h is the size of the hidden layer, and g is the final graph embedding dimension. The above mathematical expressions describe a process in which the input matrix first undergoes a single layer of convolutional operation on the graph. Through the operation $X_{bev} A_{vv}$, it captures a portion related to the input text from the global vocabulary graph. Then, it undergoes two layers of convolution, combining the words in the input text segment with their relevant words in the vocabulary graph. This process results in obtaining the global feature representation of the input text, which has dimensions of $b \times e \times g$.

After acquiring the global features G, we concatenate them with the local features X to create the graph-enriched embeddings. This combined representation is then fed into the transformer encoder module, where the features undergo layer-by-layer interactions, ensuring the retention of word order in the sentence. This approach not only preserves the sequential arrangement of words but also incorporates the background information obtained from the graph.

3.1.2 Image and Metadata Features

Images play a crucial role in fake news detection, serving as an important modality to convey information or provide supplementary context. For the images in the dataset, we employ pretrained image models to extract visual features. Before feeding the data into the model, we conduct image resizing, transforming the images to a standardized size suitable for model input.

Metadata also plays a vital role in aiding the detection of fake news. It may encompass social media metrics or categorical data, such as the number of comments, likes/dislikes, upvotes, or other ranking information. Accounts that frequently spread fake news often have fewer followers than typical users, and the posts they deliver typically have low like or upvote rates. The metadata initially needs to be standardized to a well-defined value range and then concatenated into a vector. Given the absence of a pre-defined encoder for such data, we train a lightweight multilayer perceptron (MLP) to effectively represent the input metadata.

3.2 Multimodal Feature Fusion

We employ a multilevel fusion strategy to integrate the textual, image, and metadata features obtained in the previous section. Initially, we choose to fuse the textual and image modalities, considering their common occurrence in social media, where both offer valuable information for determining the authenticity of a post. We ensure that the output dimensions of the transformer encoder and the image model are consistent. Subsequently, we explore the following four widely-used fusion strategies:

- **Concatenate**. This strategy involves merging feature tensors along a specified dimension, retaining distinct information from each modality, although it may increase the overall model parameters.
- **Add**. In this strategy, feature tensors are element-wise summed, emphasizing the complementary nature of the two modalities and mitigating their deficiencies.
- **Max Pooling**. Max pooling fusion selects the maximum value at each position from the feature tensors, highlighting dominant features from each modality and aiding in capturing the most significant information.
- **Average Pooling**. Average pooling fusion computes the average value at each position from the feature tensors of both modalities, providing a balanced representation and preserving the overall feature distribution from both sources.

Following the first-level feature fusion, we densify the fused features and proceed with the second-level fusion involving metadata features. We perform separate fusion processes because we consider metadata to play a supporting role in fake news detection. This structural choice aligns with the specific intrinsic nature of these modalities. For metadata features, we directly apply concatenate fusion to retain as much information as possible. Finally, the resulting fused features undergo dense layers and softmax activation for classification.

4 Experiments

4.1 Dataset and Benchmarks

For our experiments and investigation into multimodal fake news detection, we employ the Fakeddit dataset (Nakamura et al., 2020), which is the largest published multimodal fake news dataset to our knowledge. The Fakeddit dataset comprises one million samples spanning up to six different categories of information disorder and was collected through the pushshift API. An added advantage of this dataset is that it provides not only ground truth for binary fake/non-fake classification but also a more fine-grained distinction with 3 and 6 classes. This further categorization divides fake news into smaller categories such as "Misleading Content" or "False Connection." The dataset includes Reddit postings with comments, where many postings

Table 1: Performance Comparison on Fakeddit Dataset

Models	2-way		3-way		6-way	
	Validation	Test	Validation	Test	Validation	Test
BERT (Devlin et al., 2019)	0.9077	0.9056	0.9060	0.8997	0.8460	0.8475
Ma-RVNN (Ma et al., 2020)	0.9025	0.9036	0.9002	0.8996	0.8420	0.8453
RoBERT (Xu et al., 2022)	0.9285	0.9278	0.9214	0.9205	0.8601	0.8575
ToBERT (Xu et al., 2022)	0.9286	0.9275	0.9226	0.9216	0.8613	0.8605
Nakamura et al.	0.8929	0.8909	0.8905	0.8890	0.8600	0.8588
MVAE+ (Li et al., 2022)	0.9007	0.9003	0.8992	0.8895	0.8623	0.8612
EMAF (Li et al., 2022)	0.9246	0.9230	0.9234	0.9216	0.8856	0.8816
Kirchknopf et al.	0.9412	0.9436	0.9337	0.9326	0.9014	0.9005
AM3 (Yuan et al., 2024)	0.9310	0.9294	0.9277	0.9245	0.9065	0.9046
Ours	**0.9585**	**0.9590**	**0.9517**	**0.9522**	**0.9229**	**0.9254**

contain both text and images. Additionally, various metadata attributes are available, such as the number of upvotes and downvotes of postings, the number of comments, and a score for the post itself.

For our experiments, we selected data where all modalities are available, resulting in 560k training samples, 58k validation samples, and 58k testing samples. The standard benchmark for this task is accuracy, and since Fakeddit provides a split of the dataset into training, validation, and test sets, it facilitates directly comparable experiments. In our experiments, we focus not only on commonly compared 2-way classification tasks but also on the 3-way and 6-way tasks. This choice is driven by our intention to demonstrate that our method can delve deeper into the nuances of specific fake news types.

4.2 Experimental Setup

4.2.1 Vocabulary Graph Building

We construct our vocabulary graph using normalized point-wise mutual information (NPMI) (Bouma, 2009), defined by the equation:

$$\text{NPMI}(i,j) = -\frac{1}{\log p(i,j)} \log \frac{p(i,j)}{p(i)p(j)}$$

In this formulation, i and j represent individual words. Specifically, $p(i,j) = \frac{\#W(i,j)}{\#W}$ denotes the joint probability of words i and j appearing together, while $p(i) = \frac{\#W(i)}{\#W}$ represents the marginal probability of word i. Here, $\#W(*)$ signifies the number of sliding windows containing either a word or a pair of words, and $\#W$ is the total number of sliding windows.

To capture long-range dependencies within sentences, we set the window size larger than an entire sentence. A positive NPMI value indicates a strong semantic correlation between words, whereas a negative value signifies little to no semantic correlation. For our vocabulary graph, we establish an edge between two words if their NPMI exceeds a specified threshold. In our experiments, we set this threshold to 0.

4.2.2 Training details

Given the extensive nature of the training data, we employ Distributed Data Parallel (DDP) (Li et al., 2020) to accelerate the training process. This technique allows for data parallelism at the module level across 8 V100 GPUs. Each image is resized to 299 × 299 pixels to conform to the input dimensions required by the image model. The network is trained end-to-end using the AdamW optimizer (Loshchilov and Hutter, 2019), with a learning rate set to 1×10^{-5} and an L2 regularization term of 0.01. We conduct training over 10 epochs for each task, utilizing a batch size of 8 per GPU.

4.3 Experiment Results

To evaluate the efficacy of our proposed approach, we benchmark it against several state-of-the-art models encompassing both monomodal and multimodal architectures. The summary of these comparisons is presented in Table 1.

For monomodal models, we consider the following: (1) **BERT** (Devlin et al., 2019), fine-tuned for a straightforward classification task based on textual content, comprising the original post and comments. (2) **Ma-RVNN** (Ma et al., 2020), a

Table 2: Different Textual and Image Fusion Performance Results

Fusion Type	2-way		3-way		6-way	
	Validation	Test	Validation	Test	Validation	Test
Add	0.9543	0.9512	0.9478	0.9497	0.9206	0.9190
Max	0.9437	0.9421	0.9379	0.9394	0.9173	0.9134
Average	0.9469	0.9432	0.9402	0.9389	0.9165	0.9146
Concatenate	**0.9585**	**0.9590**	**0.9517**	**0.9522**	**0.9229**	**0.9254**

tree-structured recursive neural network designed to extract discriminative features from microblog posts by following their non-sequential propagation structure. (3) **RoBERT & ToBERT** (Xu et al., 2022), BERT-based models that concatenate the original post and associated comments into a single long text. This text is then segmented into shorter chunks for BERT-based vectorization and fed into a classifier based on an LSTM network or a transformer layer for the detection task.

For multimodal models, our selected comparisons include: (1) **BERT+ResNet50** (Nakamura et al., 2020), a straightforward model that extracts features from both the original text and images, fusing them through max pooling. (2) **MVAE+** (Li et al., 2022), which introduces a variational autoencoder to reconstruct data from multimodal representations, promoting the learning of correlations between modalities. (3) **EMAF** (Li et al., 2022), adopting an entity-centric cross-modal interaction to preserve semantic integrity and capture details of multimodal entities. (4) **Multimodal Information Disorder Model** (Kirchknopf et al., 2021), a multimodal network architecture enabling different levels and types of information fusion, including comments and metadata. (5) **AM3** (Yuan et al., 2024), featuring a novel asymmetric fusion architecture designed not only to fuse common knowledge in both modalities but also to exploit the unique information in each modality.

From the results presented in the table, it is apparent that multimodal models generally outperform their monomodal counterparts by leveraging the integration of features from both text and other dimensions. However, it is noteworthy that as the number of classification categories increases, the overall performance tends to decrease. This decline can likely be attributed to the unbalanced data distribution among the various categories and the inherent difficulty in distinguishing nuances between specific fake news categories. Despite these challenges, our proposed methods consistently achieve state-of-the-art (SOTA) performance across all classification tasks. This underscores the effectiveness of our approach in extracting and fusing modal features to enhance fake news detection.

Additionally, we have conducted experiments to investigate the impact of different fusion techniques for integrating textual and image features, as described in Section 3.2. The results, summarized in Table 2, indicate that feature concatenation yields the best performance. We believe this is due to its capacity to preserve the most comprehensive set of features from both textual and image content.

4.4 Ablation Experiments

We also conduct extensive ablation experiments to evaluate the effectiveness of each component within our model. The components assessed include local textual features (derived from BERT), global textual features (derived from VGCN), comments, image features, and metadata features. The results are presented in Table 3.

From these results, it is evident that each component of our model significantly contributes to its overall exemplary performance in fake news detection tasks. Notably, we observe that the removal of image features leads to a more substantial performance drop compared to the removal of comments or metadata. This suggests that image features play a more crucial role in debunking fake news. Furthermore, our findings indicate that local features are more critical than global features. This is likely attributable to the fact that global features capture the connections between words in the text and the overall vocabulary, but they fall short in understanding the word order and specific associations within the textual content, resulting in a decline in performance.

Table 3: Ablation Experiments Performance on Test Dataset in Fakeddit

Textual(Local)	Textual(Global)	Comment	Image	Metadata	2-way	3-way	6-way
✓	✓		✓	✓	0.9456	0.9387	0.9124
✓	✓	✓		✓	0.9316	0.9217	0.9038
✓	✓	✓	✓		0.9519	0.9479	0.9198
	✓	✓	✓	✓	0.9173	0.9094	0.8748
✓		✓	✓	✓	0.9437	0.9287	0.9089
✓	✓	✓	✓	✓	**0.9590**	**0.9522**	**0.9254**

5 Discussion

5.1 Case Analysis

We also present example cases from the Fakeddit dataset to provide a more tangible demonstration of our model's effectiveness. In Fig. 3, we showcase two examples that other baseline models incorrectly classify into different categories, where our method succeeds. Both examples contain images and main texts that offer limited information for debunking the fake news. However, by extracting both global and local textual features from the comments, our model accurately links the hidden information in the comments to the specific type of fake news. For instance, the phrase "better title" reveals that the post's title and images are misleadingly connected.

Image		
Main Text	infinity mirror the atmosphere of pluto	creepy deer
Comments	... I had a similar fake album cover before, with this title, but I can't find it anywhere... Also, this is a fun picture, as it's not really the atmosphere of a planet, but morning frost on a car roof	Better title would be creepy goat...I still had to enlarge it to see what the hell that was... I thought it was a poorly done stuffing of a bear head or something...My first thought was bear
Metadata	upvote_rate: 0.96 #comments: 4 score: 22	upvote_rate: 0.97 #comments: 7 score: 405
Groudtruth	misleading content	false connection

Figure 3: Some Cases from Fakeddit

5.2 Limitations and Further Works

As with all prior works, our methods also have certain limitations that warrant further exploration.

- Further Research on Cross-Modality Fusion. While our study extensively investigates widely-used cross-modality fusion techniques, there are still many unexplored avenues. Future research can delve deeper into enhanced cross-modality fusion methods, integrating both textual and image metadata. For example, identifying entities within images can help to better ascertain the relevance between text and visuals, which could be crucial for accurately categorizing specific types of fake news.

- Detailed Feature Representation in Visual Modality: Although our approach meticulously considers textual features, the strategy for extracting visual features relies on relatively straightforward methods using a pretrained image model. Future work could focus more on advanced visual representation techniques.

6 Conclusions

In this paper, we introduce a novel framework for detecting multimodal fake news by incorporating both global and local textual features. By leveraging the BERT model and vocabulary graph convolution networks, we effectively extract rich semantic information from both the post and its associated comments. This dual approach not only captures semantic connections within the text but also employs a global corpus representation to better align with the intricate context of social media. Subsequently, we implement a multilevel fusion technique to integrate visual and metadata information, thereby enhancing the connectivity of cross-modal features. Through extensive experimentation, we demonstrate that our method consistently achieves state-of-the-art (SOTA) performance across all classification tasks on the largest multimodal fake news dataset, Fakeddit. These results validate the effectiveness of our approach, underscoring its potential for more reliable and nuanced fake news detection.

References

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: Sentiment-aware multi-modal embedding for detecting fake news. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 41–48.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Manqing Dong, Lina Yao, Xianzhi Wang, Boualem Benatallah, Quan Z. Sheng, and Hao Huang. 2018. Dual: A deep unified attention model with latent relation representations for fake news detection. In *Web Information Systems Engineering – WISE 2018*, pages 199–209. Springer International Publishing.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. 2021. Multimodal detection of information disorder from social media. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4.

Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2022. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 24:3455–3468.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*.

Xiang Lin, Xiangwen Liao, Tong Xu, Wenjing Pian, and Kam-Fai Wong. 2019. Rumor detection with hierarchical recurrent convolutional neural network. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, pages 338–348. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: Augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval*, pages 369–382. Springer International Publishing.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2020. An attention-based rumor detection model with tree-structured recursive neural networks. *ACM Transactions on Intelligent Systems and Technology*, 11(4):42:1–42:28.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47.

Yang Xu, Jie Guo, Weidong Qiu, Zheng Huang, Enes Altuncu, and Shujun Li. 2022. "comments matter and the more the better!": Improving rumor detection with user comments. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 383–390.

Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8532–8542.