CLFFRD: Curriculum Learning and Fine-grained Fusion for Multimodal Rumor Detection

Fan Xu[1], **Lei Zeng**[1], **Bowei Zou**[2*], **AiTi Aw**[2] **and Huan Rong**[3]
[1] Jiangxi Normal University
[2] Institute for Infocomm Research, A*STAR
[3] Nanjing University of information science & Technology
{xufan, 202141600087}@jxnu.edu.cn
{zou_bowei, aaiti}@i2r.a-star.edu.sg
ronghuan@nuist.edu.cn

Abstract

In an era where rumors can propagate rapidly across social media platforms such as Twitter and Weibo, automatic rumor detection has garnered considerable attention from both academia and industry. Existing multimodal rumor detection models often overlook the intricacies of sample difficulty, e.g., text-level difficulty, image-level difficulty, and multimodal-level difficulty, as well as their order when training. Inspired by the concept of curriculum learning, we propose the Curriculum Learning and Fine-grained Fusion-driven multimodal Rumor Detection (CLFFRD) framework, which employs curriculum learning to automatically select and train samples according to their difficulty at different training stages. Furthermore, we introduce a fine-grained fusion strategy that unifies entities from text and objects from images, enhancing their semantic cohesion. We also propose a novel data augmentation method that utilizes linear interpolation between textual and visual modalities to generate diverse data. Additionally, our approach incorporates deep fusion for both intra-modality (e.g., text entities and image objects) and inter-modality (e.g., CLIP and social graph) features. Extensive experimental results demonstrate that CLFFRD outperforms state-of-the-art models on both English and Chinese benchmark datasets for rumor detection in social media.

Keywords: Rumor detection, curriculum learning, fine-grained fusion, data augmentation, sample difficulty

1. Introduction

In today's rapidly expanding digital landscape, the widespread availability of the Internet and the prevalence of social media platforms have created a global ecosystem where information spreads at unprecedented rates. Social media platforms, such as Twitter and Weibo, host vast quantities of user-generated content, much of which lacks credible validation and verification, resulting in the rapid dissemination of unverified or false information.

Automated rumor detection has become a vital necessity due to the limitations of manual methods, including low coverage and significant delays in verification. Recent years have witnessed a surge in automatic rumor detection within both academic and industrial domains (Wang et al., 2018; Zhou et al., 2020; Xu et al., 2021; Singhal et al., 2022; Zheng et al., 2022; Dhawan et al., 2022). However, current multimodal rumor detection models face three critical challenges. First, they typically select samples randomly during training without considering the complexity of samples, such as text-level, image-level, and multimodal-level difficulty. Intuitively, gradually increasing the sample difficulty during training, akin to how humans learn, could enhance rumor debunking. Second, existing multimodal rumor detection models often lack mechanisms for a deep fusion of fine-grained features from textual and visual modalities, such as entities from text and objects from images, which hinders their ability to capture intricate relations within multimodal data. Third, existing models suffer from a data scarcity issue, especially concerning images. For instance, the Twitter dataset released by Boididou et al. (2018) comprises 13,893 samples but only 514 images. Current models, whether concatenating textual and visual representations (e.g., Singhal et al., 2022; Khattar et al., 2019) or employing co-attention models (e.g., Zheng et al., 2022; Wu et al., 2021) that focus on aligning textual and visual modalities, suffer from image scarceness. Therefore, augmenting the visual modality using the textual modality presents a promising avenue to enhance the performance in the context of scarce visual data.

To address the above challenges, we introduce a framework, Curriculum Learning and Fine-grained Fusion-driven multimodal Rumor Detection (CLFFRD)[1], to automate sample selection and training by assessing sample difficulty at different training stages. Moreover, we present a fine-grained fusion strategy that combines entities from textual content with objects from images, creating a common semantic space for improved model understanding.

* Corresponding author.

[1] Code for all experiments in this paper are available at https://github.com/jxnuzl/CLFFRD

LREC-COLING 2024, pages 3314–3324
20-25 May, 2024.

In addition, we propose a novel data augmentation method utilizing linear interpolation between textual and visual modalities to generate diverse data. Furthermore, our approach incorporates deep fusion techniques, integrating intra-modality elements (such as entities from source text and image objects) and inter-modality elements (e.g., CLIP and social graph) simultaneously. Extensive experimentation validates the effectiveness of CLFFRD, demonstrating its superior performance compared to state-of-the-art methods on benchmark datasets for rumor detection in social media, encompassing both English and Chinese content.

This paper offers three major contributions.

(1) We introduce a curriculum learning strategy that adapts training according to the difficulty of samples, mirroring the natural progression of human learning.

(2) Our fine-grained fusion technique enhances multimodal model performance by merging textual and visual features, creating a shared semantic space.

(3) We present a novel data augmentation method involving linear interpolation between textual and visual modalities, enabling the generation of diverse data.

2. Related Work

Single Modality-based Approaches focus on leveraging specific features extracted from the textual modality to train a classifier. Notable aspects explored within these models include: (1) Propagation Patterns (Wu et al., 2015; Ma et al., 2017; Lao et al., 2021) have delved into the analysis of propagation patterns. For example, Wu et al. (2015) observed that rumors often start with posts by normal users, gain support from opinion leaders, and are subsequently reposted by a large number of normal users. The propagation pattern of fake news, however, differs significantly, with opinion leaders directly reposting the content, bypassing the initial stage of normal user posts. (2) User Credibility (Mukherjee and Weikum, 2015; Yuan et al., 2020; Li et al., 2019). Mukherjee and Weikum (2015) evaluated user credibility based on a combination of factors, including community engagement metrics (such as the number of answers, ratings given, comments, ratings received, disagreement, and the number of raters), inter-user agreement, typical perspective and expertise, and interaction patterns. (3) Writing Styles (Rubin et al., 2015; Potthast et al., 2018; Przybyla, 2020; Xu et al., 2020) has adopted various linguistic features, such as characters' unigrams, bigrams, and trigrams, stop words, part of speech distribution, readability values, word frequency, proportion of quoted text, and external links, as well as structural characteristics like the number of paragraphs and the average length of text.

Multimodal-based Approaches have emerged as potent tools for rumor detection, capitalizing on their ability to explore the dynamic interactions between textual and visual modalities. Dhawan et al. (2022) introduced a graph attention-based framework to incorporate the interaction between textual word information and local visual objects from images. Yuan et al. (2019) presented a graph-based rumor detection model to combine the encoding of local semantic and global structural information simultaneously. Zheng et al. (2022) proposed a GAT-based model to integrate textual, visual, and social graphs. Wang et al. (2018) introduced a GAN-based multimodal fake news detection model, to adopt the Visual Geometry Group (VGG) to extract visual features and utilize a Convolutional Neural Network (CNN) for simultaneous extraction of textual features. Recently, Xu et al. (2023) proposed a knowledge distillation driven framework to conduct incomplet modality rumor deteciton. Xu et al. (2024) presented a hierarchical graph attention networks based model to model the proposed text-image graph which can capture the different semantic interactions of the intra-modality and the inter-modality simultaneously.

However, existing rumor detection models often overlook a critical aspect: the assessment of sample difficulty at various stages of training. While Ma et al. (2022) considered the difficulty of negative samples by masking some nodes of propagation graph perspective, their model needs to discriminative negative samples in advance. In contrast to their method, we do not differentiate between negative and positive samples; instead, we introduce a holistic framework to gauge the difficulty of all samples. This framework considers three essential dimensions: text-level difficulty, image-level difficulty, and multimodal-level difficulty.

3. Methodology

3.1. Task Formulation

We define $P = \{p_1, p_2, ..., p_n\}$ as a set of posts. Each post p_i consists of $\{t_i, v_i, u_i, c_i\}$, where t_i indicates a source text, v_i denotes an image, u_i represents a user, and c_i refers to a comment. We approach rumor detection as a binary classification task, with a goal of learning a function $f(p_i) \rightarrow y$, where p_i represents the given multi-modal post, and y represents the label assigned to the post, where y=1 signifies a rumor and y=0 represents a non-rumor.

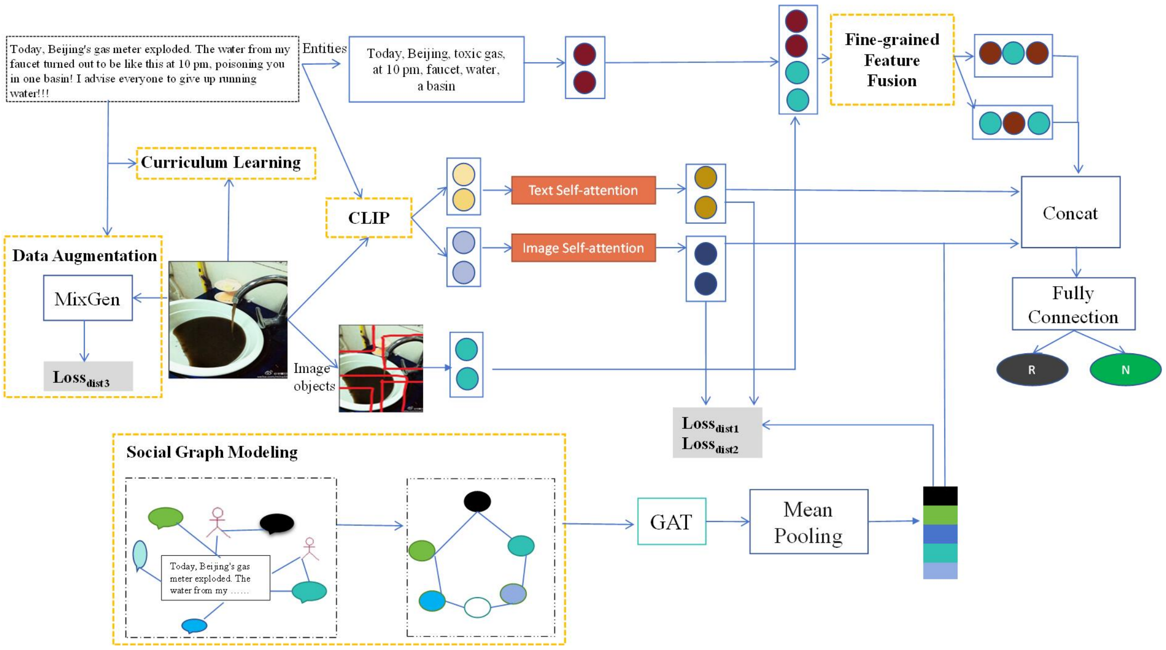

Today, Beijing's gas meter exploded. The water from my faucet turned out to be like this at 10 pm, poisoning you in one basin! I advise everyone to give up running water!!!
Entities
Today, Beijing, toxic gas, at 10 pm, faucet, water, a basin
Fine-grained Feature Fusion
Curriculum Learning
CLIP
Text Self-attention
Image Self-attention
Concat
Fully Connection
R
N
Data Augmentation
MixGen
Loss_dist3
Image objects
Loss_dist1
Loss_dist2
Social Graph Modeling
Today, Beijing's gas meter exploded. The water from my ......
GAT
Mean Pooling

Figure 1: Framework of CLFFRD; R: rumor; N: non-rumor.

3.2. Framework of CLFFRD

Figure 1 illustrates the proposed CLFFRD framework, consisting of five key modules, namely multimodal feature extraction, curriculum learning, data augmentation, fine-grained feature fusion, and the output layer. Specifically, the multimodal feature extraction module obtains representations from both intra-modality (entities from the source text and image objects) and inter-modality (CLIP and social graph). The extracted multimodal features are fed into a fine-grained feature fusion process, wherein we employ textual features to enrich image features through a linear interpolation operation, creating a synergy between textual and visual information. In the output layer module, we directly concatenate the representations from intra-modality and inter-modality, and this combined representation is further processed through a fully connected layer. To optimize our model's training process, we employ curriculum learning, involving the selection of the order of samples based on their perceived difficulty, encompassing text-level, image-level, and multimodal-level challenges. Curriculum learning helps us gradually increase the complexity of the samples during training, aligning with the natural learning progression observed in human cognition. Overall, our CLFFRD framework integrates the strengths of curriculum learning and fine-grained feature fusion to tackle the intricacies of multimodal rumor detection, enhancing the model's accuracy and effectiveness in the face of challenging rumor detection scenarios.

3.3. Multimodal Feature Extraction

3.3.1. Intra-modality Features

For a given multimodal post p_i, we adopt the large-scale pre-training model CLIP (Radford et al., 2021) to encode both textual and visual modalities, yielding F_t and F_v, respectively, as per Equations 1 and 2.

$$F_t^i = CLIP.encode\_text() \tag{1}$$

$$F_v^i = CLIP.encode\_image() \tag{2}$$

We obtain the multi-head self-attention feature of textual modality after using CLIP encoding as Equation 3.

$$Z_t^i = (||_{h=1}^{H} softmax(\frac{Q_t^i {K_t^i}^T}{\sqrt{d}})F_t^i)W_t^O \tag{3}$$

where "||" denotes concatation operation. h refers to the h-th head, $W_t^O \in R^{(d*d)}$ indicates the output of linear transformation, and d refers to the dimension size of the word embedding.

Similarly, we obtain the multi-head self-attention feature of visual modality after using CLIP encoding as Equation 4.

$$Z_v^i = (||_{h=1}^{H} softmax(\frac{Q_v^i {K_v^i}^T}{\sqrt{d}})F_v^i)W_v^O \tag{4}$$

where $W_v^O \in R^{(d*d)}$ indicates the output of linear transformation.

3.3.2. Inter-modality Social Graph Modeling

We construct a social graph $G_{social} = (V, E)$, where $V = \{post, comment, user\}$ as node. The edge is calculated by computing the cosine similarity between any pair of nodes V_i and V_j. If the value of cosine similarity as shown in equation 5 is greater than 0.5, there is an edge, otherwise there is no edge. If the cosine similarity, as defined in Equation 5, exceeds 0.5, an edge is established; otherwise, no connection is formed.

$$E_{V_i,V_j} = \frac{V_i V_j}{||V_i|| ||V_j||} \tag{5}$$

After constructing the social graph, we update the node features using a Graph Attention Network (GAT) followed by a mean pooling operation to obtain the feature of social graph as shown in Equation 6 where $MultiHead(v_i, v_j)$ and $head_h$ are defined in Equations 7 and 8, respectively.

$$F_{Social} = \sigma(\sum_{j \in N_j} MultiHead(v_i, v_j)) \tag{6}$$

$$MultiHead(v_i, v_j) = W_o ||_{h=1}^{H} (head_1, ..., head_H) \tag{7}$$

where "||" donates concatation operation, and H indicates the number of heads.

$$head_h = softmax(\frac{W_q^h V_i (W_k^h V_j)^T}{\sqrt{d}}) W_v^h V_j \tag{8}$$

where d refers to the size of dimension of node.

3.3.3. Curriculum Learning

Curriculum learning, originally proposed by Bengio et al. (2009), is a training strategy that emulates human learning proces ses, advocating the notion of commencing learning with easy samples and gradually advancing to tackle more complex ones. This strategy has consistently demonstrated its efficacy in enhancing the generalization capability and convergence rate of various models across diverse domains, including computer vision and natural language processing.

Curriculum learning aims to enhance rumor debunking through a gradual and adaptive training process. The methodology is inspired by the observed natural learning progression in human cognition, where individuals tend to start with simpler concepts before progressing to more complex ones. Our rationale for incorporating curriculum learning is rooted in its ability to guide the model through a curriculum of samples, starting from easier instances and gradually introducing more challenging ones. This mimics the learning phase observed in human cognition, where individuals are exposed to simpler concepts before tackling more intricate ones. By aligning the training curriculum with this natural learning progression, we aim to improve the model's performance by allowing it to build a robust understanding of rumors in a step-by-step manner. Drawing inspiration from the principles of curriculum learning, we introduce a novel approach by establishing three distinct scores to gauge the difficulty of samples across three scenarios: textual level, image level, and multimodal level.

Textual-level Score: To determine the difficulty of a text within a multimodal post, we employ Text-Smart Utility to generate syntax trees for all training samples, subsequently sorting the syntax tree depth values in ascending order to derive the difficulty score, $Score_{text}$. This method intuitively signifies that the deeper the syntactic tree level, the more complex the sample is.

Image-level Score: For an image within a multimodal post, we assess its difficulty by measuring entropy and sorting the entropy values of each image in ascending order to obtain the difficulty score, $Score_{image}$, as defined in Equation 9. In this context, higher information entropy corresponds to higher sample difficulty, as greater entropy values indicate richer, more intricate information content in the image.

$$Score_{image} = \sum_{i=0}^{255} P_i log P_i \tag{9}$$

where P_i represents the proportion of pixels with a grayscale value of i in an image.

Multimodal-level Score: We get $Score_{multimodal}$ by computing the cosine similarity between textual and visual features obtained from the large-scale pre-training model CLIP. Subsequently, we sort these cosine similarity values in descending order, as illustrated in Equation 10. Notably, samples with higher cosine similarity values are deemed to be comparatively easier to learn.

$$Score_{multimodal} = \frac{F_t F_v}{||F_t|| ||F_v||}. \tag{10}$$

where $F_t = CLIP.encode\_text()$, and $F_v = CLIP.encode\_image()$.

Pacing Function: The pacing function defines our course scheduling strategy, regulating the progression of training steps within each stage of the training phase. To optimize this process, we introduce a novel linear pacing function, as depicted in Equation 11, which incrementally adds training samples with each epoch. Additionally, our model explores logarithmic and exponential pacing function. Note that our experiments reveal that the linear pacing function outperforms the other two alternatives.

$$g(t) = min(1, \lambda_0 + \frac{1-\lambda_0}{T_{grow}} t) \tag{11}$$

where t denotes the specific epoch, T_{grow} indicates the maximum training epoch, and λ_0 is a hyperparameter to control the initial selected training samples.

3.3.4. Fine-grained Feature Fusion

Given a multimodal post, we begin by applying Faster-RCNN (Ren et al., 2017) to extract a set of potentially significant areas (object) from the associated image. More specifically, we extract objects from image, obtaining $R = \{r_1, r_2, r_3, ..., r_k\}$, where k is total number of objects, $r_i \in R^{D_0}$ where D_0 is the dimension size of an image, k is set to 36, and D_0 is set to 2048.

For the source text within the same multimodal post, our procedure commences with stop word filtering and entity extraction, to obtain entity set T_e. We adopt Spacy[2] and Text-smart[3] to extract entities for English and Chinese post respectively. We obtain the $T_e = \{e_1, e_2, e_3, ..., e_L\}$ as the representation of all entities, where L is total number of entities. Then, we adopt the BERT to extract textual-level features, obtaining $T'_e = \{e'_1, e'_2, e'_3, ..., e'_L\}$, $e_i \in R^{D_e}$ where D_e is the dimension size of an entity, and D_e is set to 768.

To facilitate integration, we employ an MLP to project both textual (T'_e) and visual (R) features into a common semantic space, generating T''_e and R', respectively.

Considering the varying attentional relationships between textual and visual modalities, we calculate the similarity between textual entity features and image object features to serve as weights, obtaining new entity features T'_{e-i} and object features R'_{i-e} as shown in Equations 12 and 13, respectively.

$$T'_{e-i} = \sum_{i=1}^{K} \alpha_{ij} T''_e \tag{12}$$

$$R'_{i-e} = \sum_{e=1}^{L} \alpha_{ij} R' \tag{13}$$

where $\alpha_{ij} = \frac{exp(\lambda_{1 s_{\bar{ij}}})}{\sum_{i=1}^{K} exp(\lambda_{1 s_{\bar{ij}}})}$, $s_{\bar{ij}} = [s_{ij}] + / \sqrt{\sum_{i=1}^{K} [s_{ij_+}]^2}$, and $s_{i,j} = \frac{{R'_i}^T T''_e}{\|R'_i\| \|T''_e\|}$.

Finally, we obtain the final feature representation $F_{fusion} = concat(Z_t^i, Z_v^i, T'_{e-i}, R'_{i-e}, F_{Social})$.

We perform alignment between the modeled features of the comments and the self-attention features of the text, obtaining $Loss_{dist1}$ and $Loss_{dist2}$ as shown in Equations 14 and 15, respectively.

$$Loss_{dist1} = \frac{1}{n} \sum_{i=1}^{n} (Z_t^i, F_{Social})^2 \tag{14}$$

$$Loss_{dist2} = \frac{1}{n} \sum_{i=1}^{n} (Z_v^i, F_{Social})^2 \tag{15}$$

3.3.5. Data Augmentation

Taking inspiration from the multimodal data augmentation method MixGen (Hao et al., 2023), we incorporate linear interpolation to create novel images and concatenate pairs of texts to generate fresh text samples, as delineated in Equations 16 and 17. For the newly generated text and image pairs, we ensure that similarity is maintained after augmentation. To achieve this, we utilize BERT to encode the newly generated text, T_{new} to generate $T'_{new} = BERT(T_{new})$. Simultaneously, we employ the vision transformer (Dosovitskiy et al., 2021) to encode the new generated image I_{new} to generate $I'_{new} = VIT(I_{new})$. Given that supervised contrast learning (SCL) (Khosla et al., 2017) effectively pulls together representations of the same class while segregating representations from different classes, we adopt the supervised contrast learning function to represent the data augmentation loss as Equation 18. Through this data augmentation, the model gains the capacity to acquire diverse features, thereby enriching the dataset.

$$T_{new} = concat(T_i, T_j) \tag{16}$$

$$I_{new} = \lambda I_i + (1 - \lambda) I_j \tag{17}$$

$$Loss_{dist3} = -\frac{1}{N} \sum_{i=1}^{N} log \frac{exp(\frac{sim(T'_{new}, I'_{new})}{\tau})}{\sum_j^N exp(\frac{sim(T'_{new}, I'_{new})}{\tau})} \tag{18}$$

where τ indicates a temperature parameter to control different categories.

3.3.6. Output Layer

The total loss for the rumor detection is shown in Equation 19.

$$\begin{aligned} Loss_{total} = \alpha * loss_{CE} + \beta * (Loss_{dist1} + Loss_{dist2}) \\ + \gamma * Loss_{dist3} \end{aligned} \tag{19}$$

where CE denotes cross entropy, $Loss_{CE} = ylog(\bar{y}) + (1 - \bar{y})log(1 - y)$, and $\bar{y} = softmax(MLP(F_{fusion}))$.

[2] https://spacy.io/

[3] https://ai.tencent.com/ailab/nlp/texsmart/zh/index.html

4. Experimentation

4.1. Dataset

We adopt two benchmark datasets: the English dataset Pheme (Yu et al., 2017) and the Chinese dataset Weibo (Song et al., 2019). Statistics of the datasets are presented in Table 1.

Table 1: Statistics of datasets.

	Pheme	Weibo
#Non-rumor	1,428	877
#Rumor	590	590
#Images	2,018	1,467
#Users	894	985
#Comments	7,388	4,534

4.2. Experimental Settings

Parameter Settings: The batch size is set to 64, the value of epoch is set to 64, the learning rate is set to 0.001, the optimizer is ADMA, the image size is 224*224, the patch size is set to 32, the channel size is set to 3, the total number of patches is set to 49, the size of feature dimension is set to 1024, τ in supervised contrastives learning is set to 0.05, λ_1 in equation 13 is set to 0.9, the feature extracted by using Faster-RCNN is set to 2048, and λ in equation 10 is set to 0.5. We adopt pytorch 1.10 to write our source code and execute them on a server with RTX-3090 GPU with 24 GB memory. We set the total number of head is 8. T_{grow} is set to 64, and λ_0 is set to 0.3. We adopt Adam (Kingma and Ba, 2014) to optimize our loss function.

Evaluation Metric: We employ four widely-recognized evaluation metrics: accuracy, precision, recall, and F1-Score, to assess the performance of our proposed framework and other comparative approaches.

4.3. Baselines

To assess the performance of CLFFRD, we conduct comparative studies against eight baseline models. We maintain consistent training, validation, and testing splits with these baseline systems, enabling direct comparisons of results.

EANN (Wang et al., 2018): The EANN model extracted features from text through CNN, and obtained features from images using VGG, and learned the common features between different news to obtain the invariance of events.

MVAE (Khattar et al., 2019): A multi-modal variational self encoder based model is adopted in MVAE to learn the shared representation of text and image modes.

QSAN (Tian et al., 2020): A quantum-probability based rumor detection model is proposed jointly consider the importance and stance of comments under a unified framework.

SAFE (Zhou et al., 2020): A similarity-aware multimodal model that debunks fake news from the similarity between multimodal and cross-modal features jointly.

EBGCN (Wei et al., 2021): An edge-enhanced bayesian graph convolutional networks-based model that investigates the reliability of potential relationships in propagation structures.

GLAN (Yuan et al., 2019): An integration of local semantic and global structural information-based model that debunks rumor.

MFAN (Zheng et al., 2022): A feature-enhanced attention networks-based multimodal model that combines textual, visual, and social graphs to enhances graph topology and neighborhood aggregation processes when detecting rumor.

ChatGPT[4]: A popular application showcasing the capabilities of the GPT language model is our baseline model. Since ChatGPT cannot receive image modality, we adopt the source text and the first comment as the input of ChatGPT, along with a question "judge it a rumor or not" to obtain the response, and map the results to labels (i.e.g, "yes" to rumor, and "no" to non-rumor).

4.4. Results

Model Comparison: Table 2 presents the average performance and standard deviation obtained from five executions on both the English Pheme and Chinese Weibo datasets. The results demonstrate the superiority of CLFFRD across key performance metrics, including accuracy, precision, recall, and F1-Score. This highlights the significance of curriculum learning, data augmentation, and fine-grained feature fusion in our approach. While ChatGPT has proven effective in various NLP tasks, its performance in rumor detection is not satisfactory. The insights drawn from Table 2 are as follows:

(1) Among the four multi-modal baselines (EANN, MVAE, QSAN, and SAFE), SAFE achieves the highest performance in all four measures. However, QSAN demonstrates the poorest performance across all metrics, on the Weibo dataset, which highlights the ineffectiveness of the superficial combination of textual and visual modalities in QSAN. In contrast, the EANN model, enriched by the inclusion of event information, exhibits a positive impact on rumor debunking. Notably, SAFE successfully incorporates a deep interaction between textual and visual modalities, resulting in superior performance.

(2) Among the three social graph-based baselines (EBGCN, GLAN, and MFAN), they consis-

[4] https://openai.com/blog/chatgpt

tently outperform the simpler EAN-N and MVAE models. Both EBGCN and GLAN achieve comparable performance as they incorporate structural information. However, MFAN, which combines textual, visual, and social graph-based information, outperforms the others in all metrics.

Ablation Study of Modules: Table 3 presents the performance of ablation analysis of modules, where we examine the impact of various components by considering five cases:

- **w/o CLIP**: We exclude the use of CLIP.
- **w/o Fine-grained feature fusion**: We omit the utilization of the fine-grained feature fusion.
- **w/o Data augmentation**: We disregard the inclusion of data augmentation.
- **w/o Curriculum learning**: We eliminate the application of curriculum learning, and feed the sample randomly into the proposed model.
- **w/o Social graph modeling**: We do not employ social graph module in the proposed model.

Based on the findings in Table 3, several conclusions can be drawn. 1) Curriculum learning plays a crucial role in rumor detection. The performance significantly deteriorates when this process is excluded, underscoring the importance of the sample order related to its difficulty at different training stages. 2) Data augmentation contributes to debunking rumors, as evidenced by their absence leading to a decline in performance. 3) Fine-grained feature fusion enhances the model's ability to distinguish between positive and negative samples in the corpora, which positively impacts the performance of the model. 4) The CLIP modeling is also important for multimodal rumor detection, which can create a hidden semantic space for textual and visual modalities. 5) The social graph also helps to debunk rumor, as user comments serve as valuable indicators for conducting effective rumor detection.

Ablation Study of Sample Difficulty: Table 4 presents the performance of ablation analysis of sample difficulty. Generally, the performance of multimodal-level sample difficulty performs better than textual-level and image-level difficulties, indicating the importance of the fusion between textual and visual modalities. Basically, these three sample difficulties have complementarity. We obtain promising performance improvement when combining these three sample difficulties all together. In current work, we just add up the different scores to fuse these different sample difficulties directly.

Similarity Threshold Value Setting for Social Graph Construction: Table 5 lists the performance of threshold value selection for social graph construction on Pheme and Weibo. The findings from Table 5 indicate that our model consistently achieves stable performance when the cosine similarity threshold is set to 0.5.

Table 5: Accuracy of similarity threshold value setting for social graph construction.

Cosine similarity value	Pheme	Weibo
0.1	89.64	91.82
0.2	89.93	91.88
0.3	90.05	91.95
0.4	90.08	92.06
0.5	**90.16**	**92.20**
0.6	90.09	92.15
0.7	90.01	92.18
0.8	89.95	92.01
0.9	89.69	91.89
1.0	89.43	91.57

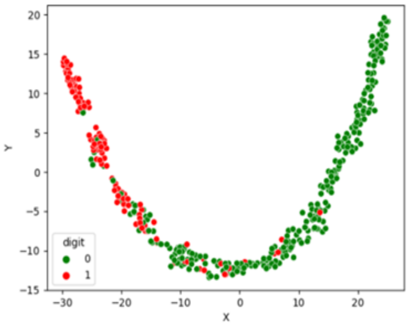


Figure 2: T-SNE visualization on Pheme.

Visualization Studies: Figures 2 and 3 display the T-SNE visualizations of the test data from Pheme and Weibo, respectively. The visualizations clearly depict the successful classification of most samples into distinct groups, demonstrating the effectiveness and strong representation capability of our proposed model.

Figures 4 shows attention visualizations for sam-

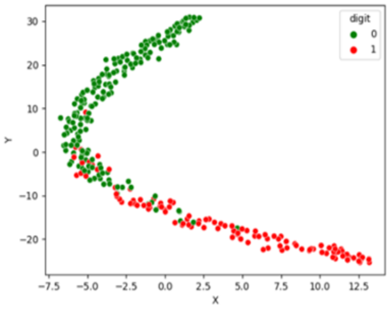


Figure 3: T-SNE visualization on Weibo.

Table 2: Performance comparison of rumor detection models on Weibo and Pheme.

	Pheme				**Weibo**			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
EANN	77.13±0.96	71.39±1.07	70.07±2.19	70.44±1.69	80.96±2.26	80.19±2.37	79.68±2.46	79.87±2.40
MVAE	77.62±0.64	73.49±0.81	72.25±0.90	72.77±0.81	71.67±0.89	70.52±0.95	70.21±1.01	70.34±0.98
QSAN	75.13±1.19	69.97±2.03	65.80±1.72	66.87±1.70	71.01±1.81	71.02±0.95	67.54±3.27	67.58±3.59
SAFE	81.49±0.84	79.88±1.22	79.50±0.81	79.68±0.70	84.95±0.85	84.98±0.82	84.95±0.91	84.96±0.86
EBGCN	82.99±0.65	81.13±0.73	79.29±0.71	79.82±0.64	83.14±2.01	85.46±2.12	81.76±1.54	81.45±1.74
GLAN	83.32±1.64	81.25±2.06	77.13±3.26	78.51±2.68	82.44±2.02	82.45±2.26	80.86±1.71	81.26±1.93
MFAN	88.73±0.83	87.07±1.41	85.51±1.65	86.16±1.04	88.95±1.43	88.91±1.60	88.13±1.68	88.33±1.53
ChatGPT	34.29±0	24.26±0	26.94±0	25.53±0	29.83±0	28.27±0	28.95±0	28.52±0
CLFFRD	**89.95±0.73**	**88.26±0.86**	**87.57±0.74**	**88.13±0.77**	**91.26±1.24**	**90.23±1.29**	**89.70±1.24**	**89.82±1.28**

Table 3: Ablation study of modules.

	Pheme				**Weibo**			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
CLFFRD	90.16	89.06	88.22	88.84	92.20	91.47	90.73	91.03
w/o Curriculum learning	88.48	88.45	88.56	88.47	90.92	91.06	90.54	90.91
w/o Data augmentation	89.15	88.24	88.15	88.22	91.18	90.60	90.24	90.48
w/o Fine-grained feature fusion	87.79	87.19	88.23	87.53	89.47	87.73	88.53	88.05
w/o CLIP	88.57	87.88	87.13	87.59	90.50	89.85	90.51	90.14
w/o Social graph modeling	88.10	87.68	87.83	87.75	90.62	89.93	90.25	89.98

ples labeled as "non-rumor" and "rumor" within the two datasets, respectively, which provides insights into the intricate interaction between textual and visual information, shedding light on how enhanced features contribute to rumor debunking. In Figure 4a, the words "police officers" and "car" highlighted in red demonstrate high attention weights and align well with specific regions in the corresponding image. However, the important word "danger" fails to align with any image regions, indicating poor alignment and predicting the sample as a rumor correctly. In contrast, in Figure 4b, the words "company", "fainted", "everyone", and "emergency assistance" can be successfully aligned with specific regions in the image. This accurate alignment contributes to the prediction of the sample as a non-rumor. These observations highlight the deep semantic interaction between the textual and visual modalities within our proposed model.

5. Conclusion

This paper presents a novel rumor detection framework that harnesses the synergistic power of curriculum learning and data augmentation. By incorporating curriculum learning, our framework intelligently orders training samples based on their different difficulties, encompassing text-level, image-level, and multimodal-level intricacies at distinct stages of training. Furthermore, our simple yet effective data augmentation method, relying on linear interpolation between textual and visual modalities, adeptly addresses the challenge of data scarceness in rumor detection. Our future endeavors will focus on the integration of user credibility into our framework for further improvement.

Acknowledgments

The authors would like to thank anonymous reviewers for their insightful comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant 62162031 and 62102187, the Natural Science Fund project in Jiangxi province under Grant 20224ACB202010, Funding of the "Jiangxi Province High level and Skilled Leading Talent Training Project", and Natural Science Foundation of Jiangsu Province (Basic Research Program) under Grant BK20210639.

Limitations

(1) Limited generalizability: Our experiments were conducted on specific datasets (Chinese Weibo and English Pheme) and may not entirely capture the intricacies of diverse rumor detection scenarios or platforms. To ensure broader applicability, future research should explore the model's generalizability to different datasets and languages.

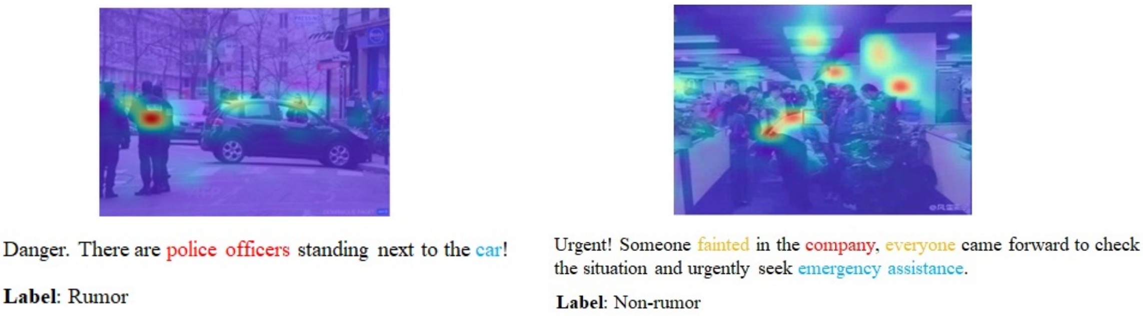
(2) Absence of real-time evaluation: Our evaluation primarily focused on offline performance metrics thus overlooking real-time or dynamic evaluation scenarios. Future work should investigate the model's performance in real-time rumor detection settings.

Ethics Statement

The benchmark datasets utilized in the project primarily reflect the cultures of the English-speaking and Chinese-speaking populace. Socio-economic biases may exist in the public and widely-used datasets, and models trained on these datasets may propagate this biases.

Table 4: Ablation study of sample difficulty.

	Pheme				Weibo			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Textual-level	88.65	88.47	88.36	88.40	89.98	89.13	88.56	89.07
Image-level	88.57	88.46	87.77	88.77	90.05	89.95	89.26	89.87
Multimodal-level	89.75	**89.26**	88.60	**89.42**	91.15	90.23	89.70	90.16
Textual-level & Image-level	89.27	88.18	**88.72**	88.63	90.74	90.30	89.83	89.97
Textual-level & Multimodal-level	90.02	89.01	88.26	88.61	91.32	90.44	90.69	90.52
Image-level & Multimodal-level	89.83	88.75	88.61	88.68	91.56	91.12	90.25	90.88
Textual-level & Image-level & Multimodal-level	**90.16**	89.06	88.22	88.84	**92.20**	**91.47**	**90.73**	**91.03**



Danger. There are police officers standing next to the car!

Label: Rumor

(a) Attention visualization on Pheme

Urgent! Someone fainted in the company, everyone came forward to check the situation and urgently seek emergency assistance.

Label: Non-rumor

(b) Attention visualization on Weibo

Figure 4: Attention visualization.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning (ICML'09)*. Montreal, Quebec.

Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, (1):71–86.

Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. Game-on: graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv 2202.12478v2*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929v2*.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. Mixgen: a new multi-modal data augmentation. *arXiv preprint arXiv:2206.08358v3*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: multimodal variational autoencoder for fake news detection. In *Proceedings of the International World Wide Web Conferences (WWW'19)*, pages 2915–2921. San Francisco, USA.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2017. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'17)*, pages 5998–6008. Online.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

An Lao, Chongyang Shi, and Yayi Yang. 2021. Rumor detection with field of linear and non-linear propagation. In *Proceedings of the Web Conference (WWW'21)*, pages 3178–3187. Ljubljana, Slovenia.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. pages 1173–1179. Florence, Italy.

Jiachen Ma, Yong Liu, Meng Liu, and Meng Han. 2022. Curriculum contrastive learning for fake news detection. In *Proceedings of the 31th ACM International Conference on Information and Knowledge Management (CIKM'22)*, pages 4309–4313. Atlanta, USA.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 708–717. Vancouver, Canada.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the ACM International Conference on Infomlation and Knowledge Management (CIKM'15)*, pages 353–362. Melbourne, Australia.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 231–240. Melbourne, Australia.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the thirty-fourth AAAI Conference on Artificial Intelligence (AAAI'20)*, pages 490–497. New York, USA.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*. Online.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Victoria Rubin, Niall Conroy, and Yimin Chen. 2015. Towards news verification: deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, pages 1–11. Hawaii, USA.

Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Proceedings of the Web Conference (WWW'22)*, pages 726–734. Online.

Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.

Tian Tian, Yudong Liu, Xiaoyu Yang, Yuefei Lyu, Xi Zhang, and Binxing Fang. 2020. Qsan: a quantum-probability based signed attention network for explainable false information detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20)*, pages 1445–1454. Online.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'18)*, pages 849–857. London, UK.

Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: edge-enhanced bayesian graph convolutional networks for rumor detection. *arXiv preprint arXiv:2107.11934*.

Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE'15)*, pages 651–662. Seoul, Korea.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Proceedings of the Findings of the Association for Computational Linguistics (ACL-IJCNLP'21)*, pages 2560–2569. Online.

Fan Xu, Pinyun Fu, Qi Huang, Bowei Zou, AiTi Aw, and Wang Mingwen. 2023. Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 13492–13503. Singapore.

Fan Xu, S. Sheng Victor, and Mingwen Wang. 2020. Near real-time topic-driven rumor detection in source microblogs. *Knowledge-Based Systems*, 207(106391):1–9.

Fan Xu, S. Sheng Victor, and Mingwen Wang. 2021. A unified perspective for disinformation detection and truth discovery in social sensing: a survey. *ACM Computing Surveys*, 55(1):1–33.

Fan Xu, Lei Zeng, Qi Huang, Keyu Yan, Mingwen Wang, and S. Sheng Victor. 2024. Hierarchical graph attention networks for multi-modal rumor detection on social media. *Neurocomputing*, 569(127112):1–11.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*, pages 3901–3907. Melbourne, Australia.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *Proceedings of the 19th IEEE International Conference on Data Mining (ICDM'19)*, pages 796–805. Beijing, China.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING'20)*, pages 5444–5454. Online.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, pages 2413–2419. Messe Wien, Vienna, Austria.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: similarity-aware multi-modal fake news dectection. *arXiv preprint arXiv:2003.04981*.