LREC-COLING 2024

Interpretable Short Video Rumor Detection based on Modality Tampering

Kaixuan Wu[1], Yanghao Lin[1], Donglin Cao[1]*, Dazhen Lin[1]

[1]School of Informatics, Xiamen University, China

May 2024

Introduction

Motivation

1. With the rise of short video platforms, rumors have also exploded in the video formats.
2. Most of the existing multimodal rumor detection work focuses on the fusion of different modal features, ignoring some characteristics of short rumor videos:
   a) rumor videos contain inconsistent information among different modalities.
   b) suffer from serious information tampering, such as manipulating textual content and splicing irrelevant image and audio information.
3. Lack of interpretability of existing methods.

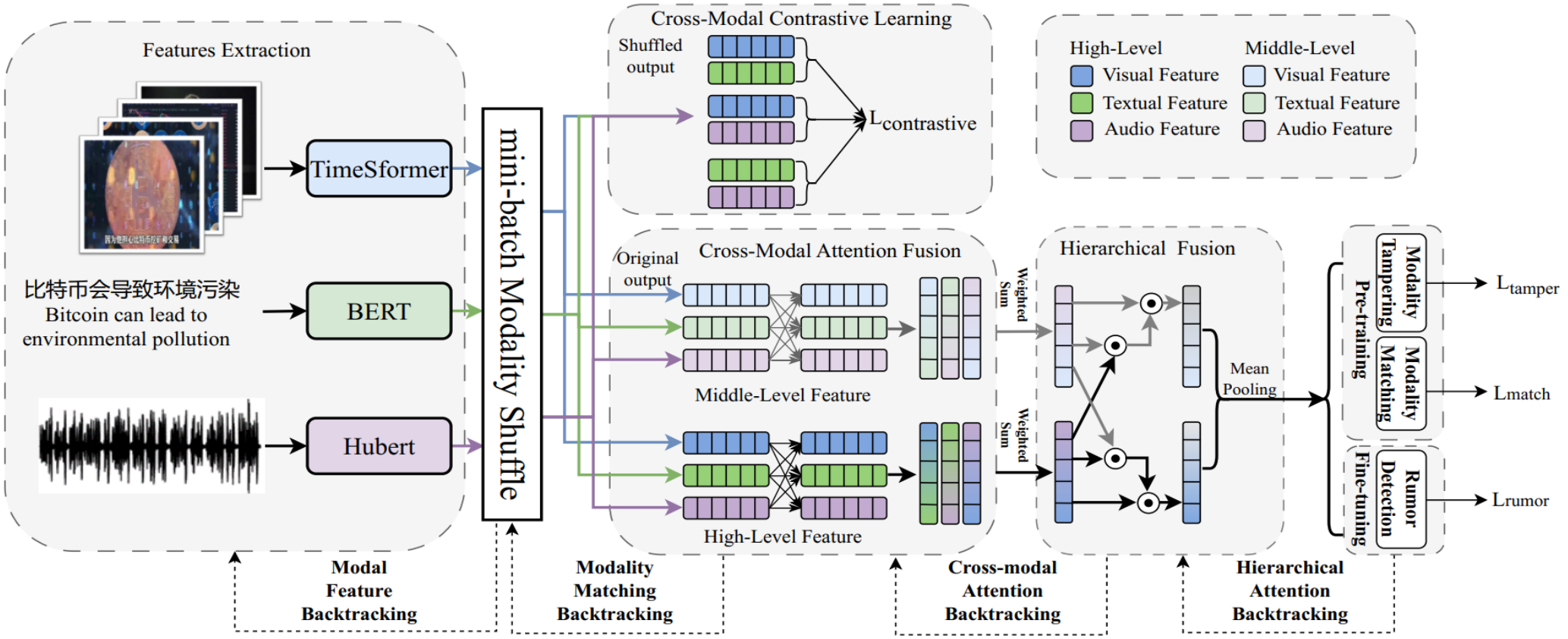
Introduction

Contributions

1. Aiming at the problem of deliberate tampering in short videos, we propose a short video rumor detection method based on modality tampering.
2. We extended a short video rumor dataset and constructed the tampering dataset to support the task of modality tampering detection.
3. We use the attention-backtracking mechanism to find local features that may have been tampered with to explain whether the short video is a rumor.
4. We conducted extensive ablation experiments to demonstrate the effectiveness of the proposed method.

Method

Model Architecture Overview of SVRPM



- Feature extraction: Extract visual, textual, and audio features.
- Mini-batch Modality Shuffle: Randomly shuffle features.

- Cross-modal Contrastive Learning
- Cross-modal & Hierarchical Fusion
- Modality Tampering Backtracking

Method

Multimodal Feature Extraction

- Textual Feature Extraction

$$H_T = BERT(w_1, w_2, \dots, w_n)_{[\mathrm{CLS}]}$$

- Visual Feature Extraction

$$H_V = TimeSformer(f_1, f_2, \dots, f_n)_{[\mathrm{CLS}]}$$

- Audio Feature Extraction

$$\{h_1, h_2, \dots, h_n\} = Hubert(a_1, a_2, \dots, a_n)$$

$$H_A = \frac{1}{n}\sum_{i=1}^{n} h_i$$

Cross-Modal Fusion

As an example, we introduce cross-modal fusion with high-level features.

$$M_{TVA} = [H_T, H_V, H_A]^T$$

$$\alpha = Softmax(Attention(q_i, M_{TVA}))$$

$$H_{TVA_H} = Sum(\alpha * M_{TVA})$$

Similarly

We can obtain the middle-level feature H_{TVA_M}

Hierarchical Fusion

$$M_{HM} = [H_{TVA_H}, H_{TVA_M}]^T$$

$$\alpha_H = Softmax(Attention(q_H, M_{HM}))$$

$$\alpha_M = Softmax(Attention(q_M, M_{HM}))$$

$$\widehat{H}_{TVA_H} = Sum(\alpha_H * M_{HM})$$

$$\widehat{H}_{TVA_M} = Sum(\alpha_M * M_{HM})$$

$$H_{TVA_{HF}} = MeanPooling(\begin{bmatrix} \widehat{H}_{TVA_H} \\ \widehat{H}_{TVA_M} \end{bmatrix})$$

Method

Tampering Selector

Tampering with some words of the video title.

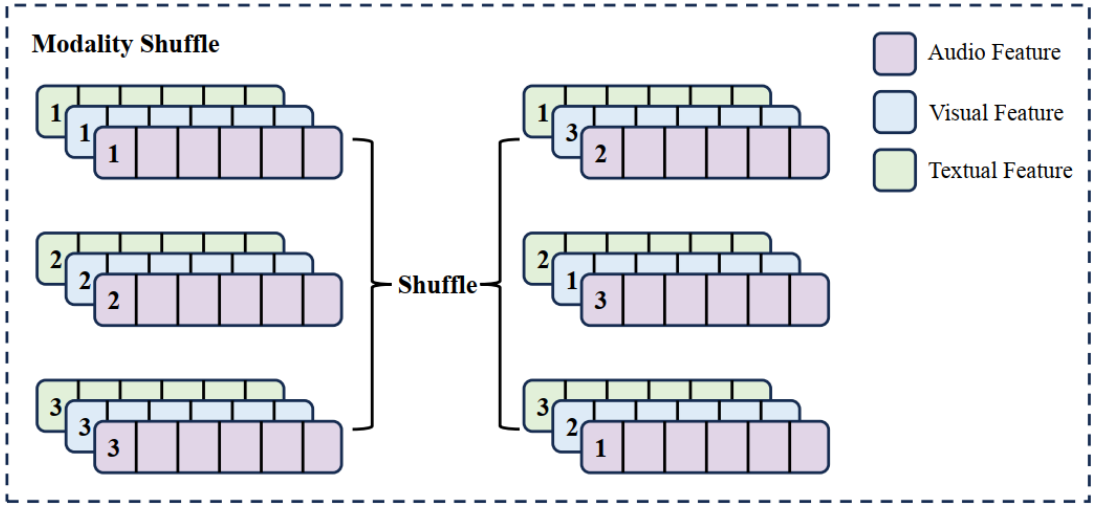
1). Use “HanLP”[1] to perform lexical segmentation, then take all nouns, adjectives, and adverbs as tampered words for each video title.

2). Use “Synonyms”[2] to obtain the four words with the closest semantic similarity to the tampered words as candidate words.

3). For each video title, 1 to 3 words to be tampered.

Before inputting the text into the model, tampering selector determines with a certain probability whether or not to tamper.

Modality Shuffle



Different colors represent different modalities, and different numbers represent different samples in a mini-batch.

1. https://github.com/hankcs/HanLP
2. https://github.com/chatopera/Synonyms

Method

Loss Function

Cosine Similarity

$$Sim_{TV} = \frac{H_T H_V^T}{\|H_T\|\|H_V\|}$$

$$Sim_{TA} = \frac{H_T H_A^T}{\|H_T\|\|H_A\|}$$

$$Sim_{VA} = \frac{H_V H_A^T}{\|H_V\|\|H_A\|}$$

Contrastive Loss of Modality Tampering

$$L_{c_t} = \sum_{i=1}^{N} (-1)^{m_i} (Sim_{T_i V_i} + Sim_{T_i A_i}) \quad \text{Eq.1}$$

N is the number of samples in a mini-batch

$$L_{tamper} = -\sum_{i=1}^{N} \left[m_i \log\left(F_t(H_i)\right) + (1 - m_i) \log\left(F_t(H_i)\right) \right] \quad \text{Eq.2}$$

$$L_t = L_{c_t} + L_{tamper} \quad \text{Eq.3}$$

F_t, F_m denote the linear layer

$$L_{c_m} = \sum_{i=1}^{N} (-1)^{n_i} (Sim_{T_i V_i} + Sim_{T_i A_i} + Sim_{V_i A_i}) \quad \text{Eq.4}$$

$$L_{match} = -\sum_{i=1}^{N} \left[n_i \log\left(F_m(H_i)\right) + (1 - n_i) \log\left(F_m(H_i)\right) \right] \quad \text{Eq.5}$$

$$L_m = L_{c_m} + L_{match} \quad \text{Eq.6}$$

Method

Modality Tampering Backtracking

Algorithm 1 Modality Tampering Backtracking

Input: $H_{TVA_M}, H_{TVA_H}, H_T, H_V, H_A, Token$
Output: Modality local original features.

Initialize: $Attn_{Hr}, Attn_{CMH}, Attn_{CMM}$
$Attn_{xMH}, Attn_{xMM} \leftarrow \boldsymbol{x}$ modal high-level and middle-level attention, $\boldsymbol{x} \in \{T, V, A\}$
$\left[Score_H, Score_M\right]^T = Attn_{Hr}(H_{TVA_H}, \left[H_{TVA_H}, H_{TVA_M}\right]^T) + Attn_{Hr}(H_{TVA_M}, \left[H_{TVA_H}, H_{TVA_M}\right]^T)$, (Equation 7, 8)
if $Score_H > Score_M$ **then**
  $\left[Score_T, Score_V, Score_A\right]^T = Sum\left(Attn_{CMH}\left(\left[H_T, H_V, H_A\right]^T, \left[H_T, H_V, H_A\right]^T\right)\right)$, (Equation 4)
  modality $\boldsymbol{x} = Max_{modal}\left(\left[Score_T, Score_V, Score_A\right]^T\right)$, the values of $\boldsymbol{x}$ are T, V, A
  $\left[Score_{Token_1}, \cdots, Score_{Token_n}\right]^T = Attn_{xMH}\left(Token_{cls}, \left[Token_1, \cdots, Token_n\right]^T\right)$, Token from modality $\boldsymbol{x}$
else
  $\left[Score_T, Score_V, Score_A\right]^T = Sum\left(Attn_{CMM}\left(\left[H_T, H_V, H_A\right]^T, \left[H_T, H_V, H_A\right]^T\right)\right)$, (Equation 4)
  modality $\boldsymbol{x} = Max_{modal}\left(\left[Score_T, Score_V, Score_A\right]^T\right)$, the values of $\boldsymbol{x}$ are T, V, A
  $\left[Score_{Token_1}, \cdots, Score_{Token_n}\right]^T = Atten_{xMM}\left(Token_{cls}, \left[Token_1, \cdots, Token_n\right]^T\right)$, Token from modality $\boldsymbol{x}$
end if
Local feature index: $Index = Max_{index}\left(\left[Score_{Token_1}, \cdots, Score_{Token_n}\right]^T, k\right)$
return Local feature: $Feature_x\,(index)$ =0

Experiments

Dataset

Dataset	Split	Rumor	non-Rumor	Total
Ours	train	467	4795	5262
	test	117	1204	1321
FakeSV[1]	train	1233	1303	2536
	test	304	238	542

Evaluation Metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

1. Qi, Peng, et al. "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 12. 2023.

Results

Results of pre-training tasks

Task	Acc	P	R	F1
Modality Matching	82.56	84.96	79.05	81.90
Modality Tampering	86.41	88.82	83.32	85.98

Experimental results(%) for the modality matching task and the modality tampering task.

Compare with Baselines

Method	Ours						FakeSV					
	Acc	F1	P_0	P_1	R_0	R_1	Acc	F1	P_0	P_1	R_0	R_1
SAFE(2020)	96.90	90.98	78.36	98.99	89.74	97.59	77.49	77.43	84.21	71.01	73.68	82.35
ViLT*(2021)	81.53	65.46	29.04	97.15	75.21	82.14	–	–	–	–	–	–
VideoMae(2022)	98.03	93.55	94.17	98.36	82.91	99.50	73.80	72.99	74.40	72.86	81.25	64.29
MEA(2022)	94.70	86.19	64.07	99.13	91.45	95.02	70.66	70.51	**86.43**	61.52	56.58	**88.66**
CHEF(2022)	97.58	92.67	84.55	98.91	88.89	98.42	–	–	–	–	–	–
SVRPM(ours)	**99.39**	**98.15**	**95.04**	**99.83**	**98.29**	**99.50**	**79.34**	**78.55**	78.07	**81.50**	**87.83**	68.49

Results(%) of different methods on our short video rumor dataset and FakeSV dataset. "*" denotes that the text content is in English. The subscript "0" represents "Rumor as Positive" and "1" denotes "non-Rumor as Positive" in computing the precision, and recall. "F1" denotes macro-F1 values. "–" means that corresponding experiments were not carried out due to lack of partial data. The best performance is highlighted in boldface.

Results

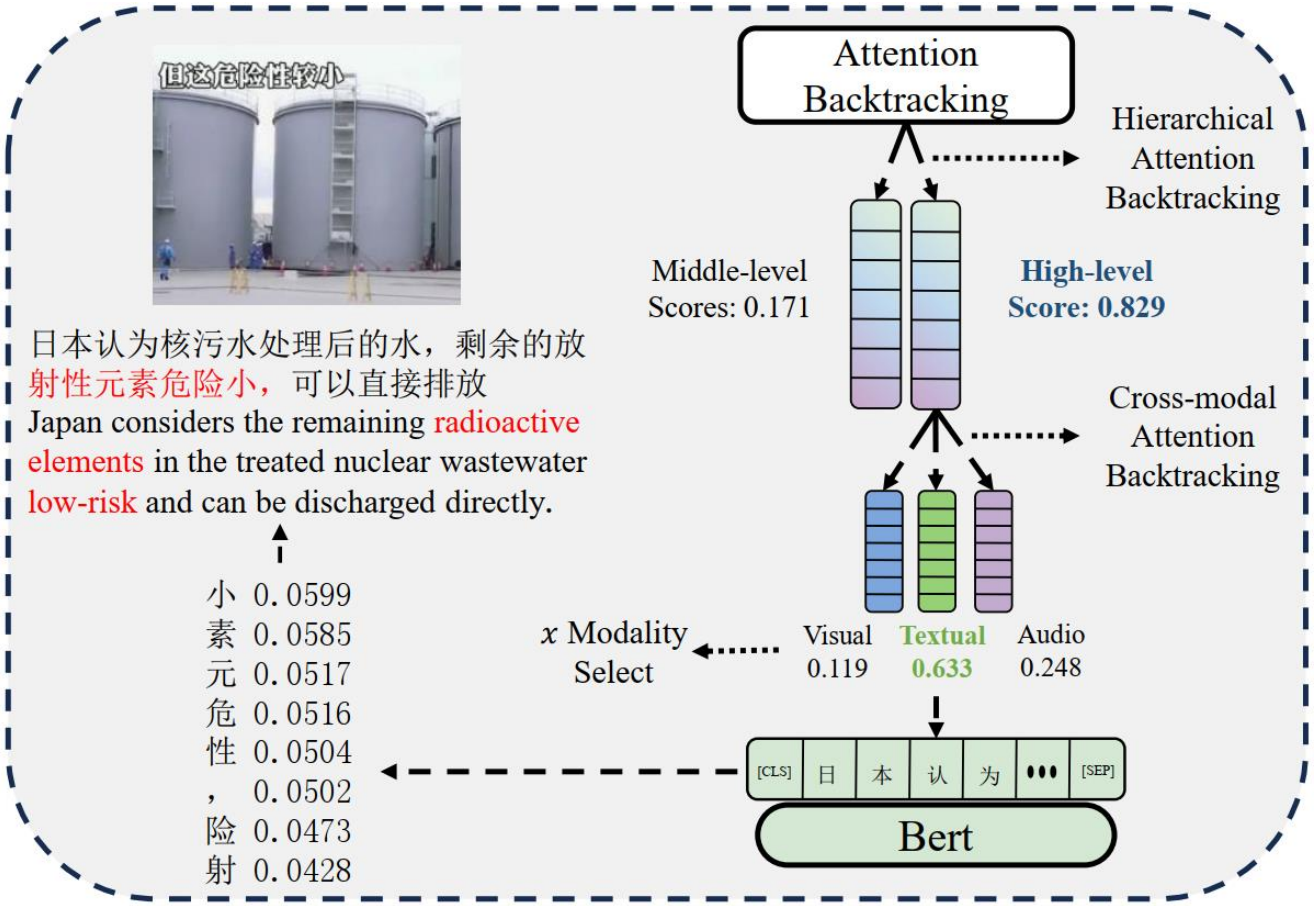
Ablation Results

Method	Modality	CMF	HF	Tamper	Match	Acc	P	R	F1
Concat	T, V	✗	✗	✗	✗	93.19	78.16	92.02	83.18
Concat	T, V, A	✗	✗	✗	✗	96.37	85.82	96.46	90.23
SVRPM	T, V	✓	✗	✗	✗	98.18	95.33	93.22	94.24
SVRPM	T, V	✗	✓	✗	✗	98.18	94.37	94.37	94.37
SVRPM	T, V, A	✓	✗	✗	✗	98.86	97.00	95.90	96.44
SVRPM	T, V, A	✗	✓	✗	✗	99.17	**97.60**	97.23	97.41
SVRPM	T, V, A	✓	✓	✓	✗	99.24	97.31	98.04	97.67
SVRPM	T, V, A	✓	✓	✗	✓	99.24	96.36	**99.20**	97.73
SVRPM	T, V, A	✓	✓	✓	✓	**99.39**	97.44	98.90	**98.15**

Performance(%) of ablation experiments. For simplicity, modalities are abbreviated("T": Textual modality, "V": Visual modality, "A": Audio modality). "CMF" denotes cross-modal fusion, "HF" denotes hierarchical fusion, and Tamper, Match stand for pre-training tasks, respectively

Results

Visualization of Modality Tampering Backtracking



Through this backtracking analysis, peak attention aligns with the item most pertinent to the query.

Conclusion

- Design an interpretable short video rumor detection model based on modality Tampering.
- Extended a short video rumor dataset and constructed the tampering dataset.
- Visualize the local tampering features using the modality tampering backtracking to improve the interpretability.