Harmfully Manipulated Images Matter in Multimodal Misinformation Detection

Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, Ximing Li**

ACM Multimedia 2024
Melbourne, Australia

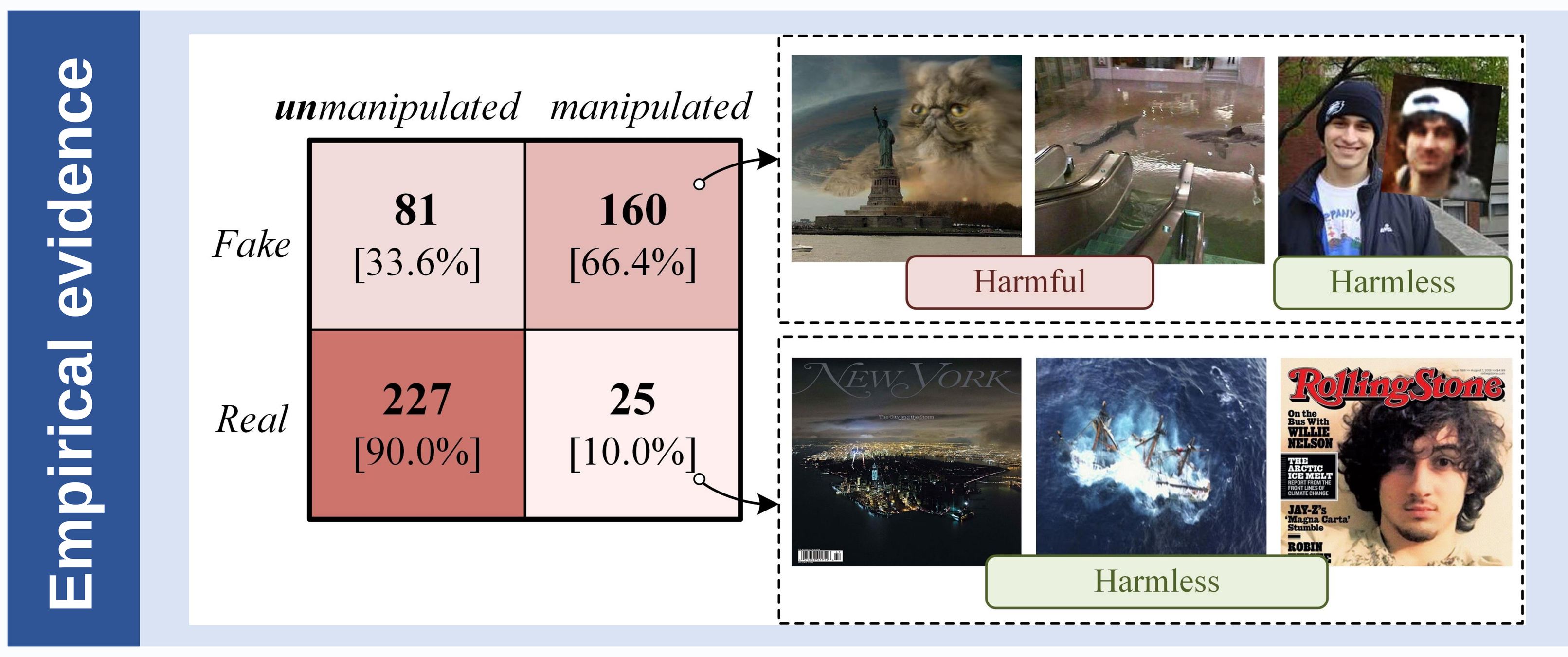
Motivation

➢ **Manipulation is a crucial feature in multimodal misinformation detection**

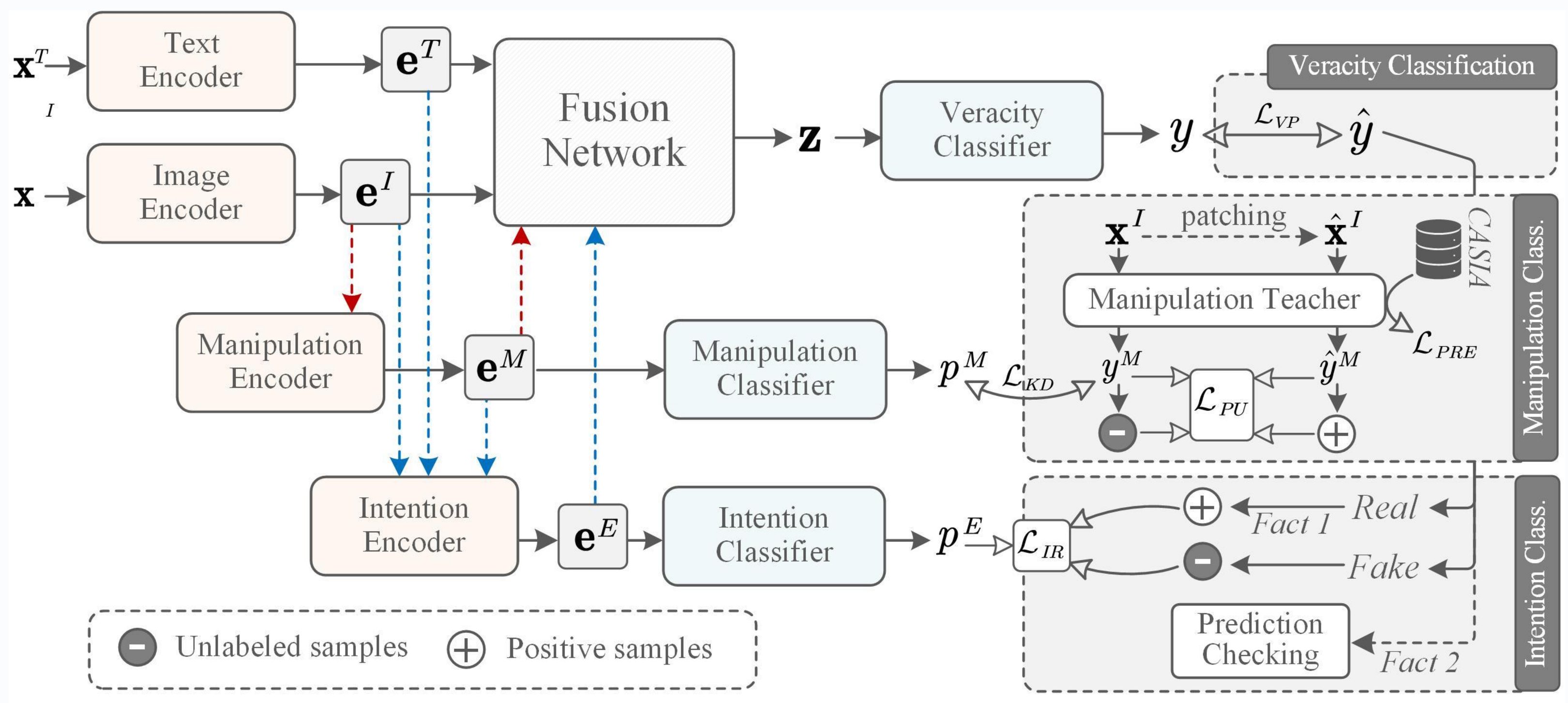
As surveys, a majority of fake articles may contain **manipulated images** created by various techniques.

➢ **Not all manipulated information is misinformation**

The images of fake articles are more likely with harmful intentions, e.g., deception, but the ones of real articles are with harmless intentions, e.g., watermarking.



Our Method: HAMI-M3D



We propose to detect misinformation by extracting distinctive **manipulation features** that reveal whether the image is manipulated, as well as **intent features** that differentiate between harmful and harmless intentions behind the manipulation.

Module 1: Feature Encoders Module

This module consists of four specific feature encoders, including text encoder, image encoder, manipulation encoder, and intention encoder.

Module 2: Feature Fusion Module

Given these extracted features, the feature fusion module utilizes a multi-head attention network to integrate them into one fused feature.

Module 3: Predictors Module

This module contains three predictors trained on three different tasks:

➢ **Task A: Veracity Classification**

Utilizing the fused feature, a linear veracity classifier is employed to predict the veracity label.

➢ **Task B: Manipulation Classification**

✓ **Knowledge distillation**: Training a manipulation teacher and then distilling its predictions to the manipulation classifier.

✓ **Pre-training**: Pre-training the manipulation teacher with a benchmark dataset on image manipulation detection, e.g., CASIAv2.

✓ **Positive and Unlabeled (PU) learning**: Given an image, we generate its manipulated version, and it is naturally assigned as "manipulated", the other is "unlabeled".

➢ **Task C: Intention Classification with PU learning**

✓ If the image of the real article is manipulated, its intention must be harmless; But if the image of the fake article is manipulated, its intention may be harmful or harmless.

✓ If the image of one article is manipulated by a harmful intention, the veracity label of this article must be fake; But if the image of one article is manipulated by a harmless intention, its veracity label may be real or fake.

Experimental Results

Method	Accuracy	Macro F1	Real			Fake			Avg. Δ
			Precision	Recall	F1	Precision	Recall	F1	
Dataset: *GossipCop*									
Basic model	87.77±0.56	79.51±0.44	91.55±0.41	93.36±1.20	92.37±0.41	69.96±1.10	63.30±1.46	66.92±0.58	-
Basic model + HAMI-M³D	88.45±0.20*	80.32±0.43*	91.93±0.14	94.08±0.33*	92.83±0.12	71.99±0.80*	64.59±0.74*	67.63±0.78*	**+0.90**
SAFE [56]	87.78±0.31	79.22±0.49	91.22±0.30	93.34±0.47	92.37±0.20	70.66±1.32	63.12±1.50	66.66±0.84	-
SAFE + HAMI-M³D	88.53±0.24*	79.87±0.30*	91.90±0.31*	94.32±0.54*	92.95±0.20*	72.19±1.30*	64.44±0.73*	67.88±0.51*	**+0.96**
MCAN [48]	87.66±0.59	78.89±0.34	90.89±0.78	94.07±1.27	92.19±0.46	71.01±1.09	60.37±1.21	65.29±0.87	-
MCAN + HAMI-M³D	88.27±0.57*	79.87±0.36*	91.72±0.35*	95.13±1.21*	93.05±0.41*	72.69±0.96*	62.64±1.21*	66.65±0.32*	**+1.21**
CAFE [8]	87.40±0.71	79.51±0.61	91.07±0.25	93.84±1.28	92.16±0.50	71.60±1.39	61.16±1.10	66.24±0.72	-
CAFE + HAMI-M³D	88.18±0.44*	80.43±0.48*	91.50±0.45	94.46±1.00*	92.80±0.31*	72.84±0.83*	62.51±0.90*	67.58±0.83*	**+0.91**
BMR [49]	87.26±0.46	79.03±0.64	90.89±0.24	93.99±0.59	92.14±0.29	71.15±1.23	60.37±1.21	65.51±1.01	-
BMR + HAMI-M³D	87.95±0.27*	79.99±0.57*	91.40±0.51*	94.73±0.75*	93.14±0.19*	72.26±0.73*	62.94±0.89*	66.80±1.09*	**+1.11**
Dataset: *Weibo*									
Basic model	90.87±0.34	90.75±0.34	91.08±0.23	90.17±0.85	90.62±0.40	90.87±0.70	91.41±0.28	91.29±0.29	-
Basic model + HAMI-M³D	91.62±0.66*	91.61±0.66*	91.83±0.87*	93.23±0.56*	91.39±0.76*	92.52±0.89*	91.87±0.64	91.84±0.62*	**+1.11**
SAFE [56]	91.06±0.88	91.04±0.89	91.09±1.25	90.51±0.90	90.73±1.04	91.27±0.78	91.57±1.14	91.36±0.85	-
SAFE + HAMI-M³D	92.22±0.91*	92.22±0.93*	91.15±1.08	94.22±0.84*	92.14±0.92*	94.34±1.00*	91.34±1.09	92.30±0.66*	**+1.42**
MCAN [48]	90.99±0.83	90.99±0.83	89.66±0.82	92.24±1.10	90.81±0.90	92.69±0.80	89.92±0.99	91.20±0.79	-
MCAN + HAMI-M³D	92.01±0.80*	92.01±0.80*	90.44±0.70*	93.37±0.87*	91.88±0.85*	93.59±0.74*	90.84±0.78*	92.17±0.76*	**+0.98**
CAFE [8]	90.99±0.78	90.98±0.78	90.31±0.72	91.19±1.09	90.73±0.97	91.70±1.26	90.81±1.03	91.24±0.60	-
CAFE + HAMI-M³D	91.95±1.06*	91.84±1.01*	91.25±0.55*	92.38±1.04*	91.66±0.91*	92.99±0.83*	91.93±0.91*	92.11±0.75*	**+1.02**
BMR [49]	90.17±0.92	90.15±0.93	90.09±1.20	89.60±0.85	89.81±1.00	90.36±0.93	90.71±0.78	90.50±0.81	-
BMR + HAMI-M³D	91.74±0.40*	91.68±0.40*	91.01±0.92*	93.17±0.82*	91.56±0.43*	93.40±0.84*	91.29±0.67*	91.81±0.38*	**+1.79**
Dataset: *Twitter*									
Basic model	65.08±1.18	63.91±1.09	57.29±1.26	66.67±1.01	61.48±1.56	72.04±0.96	62.41±0.92	65.35±1.01	-
Basic model + HAMI-M³D	66.27±0.66*	65.67±1.27*	59.70±1.16*	69.70±0.71*	62.46±1.08*	73.19±0.93*	64.12±1.12*	67.86±0.82*	**+1.84**
SAFE [56]	66.43±0.33	66.33±0.32	58.28±0.50	73.63±1.38	64.47±0.53	74.94±0.84	61.78±1.26	68.34±0.69	-
SAFE + HAMI-M³D	67.15±0.96*	67.00±0.89*	59.32±0.90*	74.05±0.99	65.65±0.70*	76.49±0.60*	63.58±1.09*	68.77±0.94	**+0.98**
MCAN [48]	65.82±0.64	65.24±1.34	58.30±1.07	63.66±1.03	61.16±1.23	71.70±1.03	67.42±1.39	69.33±1.22	-
MCAN + HAMI-M³D	67.14±1.11*	66.58±1.21*	60.63±0.99*	64.94±1.04*	62.55±1.28*	72.86±0.82*	68.77±1.12*	70.61±1.10*	**+1.43**
CAFE [8]	65.62±0.58	65.04±0.48	58.39±0.90	66.24±1.48	62.05±0.21	72.37±1.28	65.16±1.06	68.57±1.05	-
CAFE + HAMI-M³D	65.89±1.30	65.37±0.87	59.91±0.55*	67.28±1.17*	63.60±0.64*	73.42±1.18*	68.76±1.12*	70.49±1.06*	**+1.41**
BMR [49]	67.12±0.74	66.64±1.28	59.09±0.61	72.62±1.28	64.43±1.28	75.10±1.13	62.56±0.91	68.65±1.17	-
BMR + HAMI-M³D	67.84±0.83*	67.68±0.82*	60.01±0.88*	73.31±1.28*	65.65±0.92*	76.27±1.03*	64.32±0.98*	69.71±0.91*	**+1.08**

➢ HAMI-M3D outperform 5 baselines across 3 datasets

Key Takeaways

➢ **Findings:** We introduce image manipulation features into multimodal misinformation detection, and find that only the article that has been manipulated by harmful intention is misinformation.

➢ **Method**: We primarily propose three tasks: veracity classification, manipulation classification, and intention classification, which respectively detect whether an article is misinformation, whether its image has been manipulated, and whether this manipulation is harmful.

➢ **Experiments**: By comparing with baseline models, we have demonstrated the effectiveness of our model.

My homepage

Github Repo