Meme Trojan: Backdoor Attacks Against Hateful Meme Detection via Cross-Modal Triggers

Ruofei Wang[1,2], Hongzhan Lin[1], Ziyuan Luo[1,2], Ka Chun Cheung[2], Simon See[2], Jing Ma[1], Renjie Wan[1*]

[1]Department of Computer Science, Hong Kong Baptist University
[2]NVIDIA AI Technology Center, NVIDIA
{ruofei, danielhzlin, ziyuanluo}@life.hkbu.edu.hk, {chcheung, ssee}@nvidia.com, {majing, renjiewan}@hkbu.edu.hk

Abstract

Hateful meme detection aims to prevent the proliferation of hateful memes on various social media platforms. Considering its impact on social environments, this paper introduces a previously ignored but significant threat to hateful meme detection: backdoor attacks. By injecting specific triggers into meme samples, backdoor attackers can manipulate the detector to output their desired outcomes. To explore this, we propose the ***Meme Trojan*** framework to initiate backdoor attacks on hateful meme detection. Meme Trojan involves creating a novel Cross-Modal Trigger (CMT) and a learnable trigger augmentor to enhance the trigger pattern according to each input sample. Due to the cross-modal property, the proposed CMT can effectively initiate backdoor attacks on hateful meme detectors under an automatic application scenario. Additionally, the injection position and size of our triggers are adaptive to the texts contained in the meme, which ensures that the trigger is seamlessly integrated with the meme content. Our approach outperforms the state-of-the-art backdoor attack methods, showing significant improvements in effectiveness and stealthiness. We believe that this paper will draw more attention to the potential threat posed by backdoor attacks on hateful meme detection.

Introduction

With the rise of social media platforms (*e.g.*, Twitter, Reddit, *etc*.), memes, a kind of multimodal content, have emerged as popular mediums to express users' ideas and emotions (Kiela et al. 2020). As memes may convey hateful and satirical messages, leading to online abuse and hate speech (Vickery 2014; Kiela et al. 2020) (see Fig. 1 (I)), hateful meme detection is proposed to mitigate these societal risks. Despite the significant achievement in hateful meme detection (Zhu, Lee, and Chong 2022; Koutlis, Schinas, and Papadopoulos 2023; Lin et al. 2024a), Aggarwal *et al*. (Aggarwal et al. 2023) have revealed that simple adversarial examples can deceive the hateful meme detector at the inference phase. This investigation uncovers the potential security risk associated with hateful meme detection and underscores the urgent need for further exploration.

During the training stage of hateful meme detectors, a realistic threat is caused by **backdoor attacks** (Li et al. 2022). Such a risk usually arises from the use of third-party datasets that may contain poisoned samples (Gu, Dolan-Gavitt, and Garg 2017) and is significantly difficult to detect (Liu et al. 2020). Generally, attackers can inject a backdoor into the victim model by poisoning the training data, thereby manipulating the model's behavior during the inference. As shown in the Fig. 1 (II), the victim model correctly classifies the benign samples (1_{st} row: without triggers) while giving malicious results when encountering poisoned memes (2_{nd} row: with triggers). This attack enables malicious users to bypass hateful meme detectors, facilitating the dissemination of hateful memes. However, the corresponding exploration of such an attack still leaves a blank.

Memes are formed by an image and a short piece of text embedded within it (Kiela et al. 2020), showing a unique characteristic that text coexists with the image (Koutlis, Schinas, and Papadopoulos 2023). Such a characteristic and the automatic detection pipeline make current backdoor attack methods designed for uni-modality (*i.e.*, image (Gu, Dolan-Gavitt, and Garg 2017; Liu et al. 2020; Li et al. 2021) or text (Chen et al. 2021; Qi et al. 2021b)) invalid. First, existing backdoor attacks (Chen et al. 2021; Walmer et al. 2022) designed for text modality necessitate the prior acquisition of the text component to inject triggers. However, the text information is inaccessible for humans in an automatic detection system, resulting in **low effectiveness**. Second, if a malicious user inputs texts and injects triggers manually, the poisoned texts show inconsistency with the original texts embedded in the image, **reducing stealthiness**. As shown in case (a) of Fig. 1 (II), the extra word "Consider" appears extremely doubtful, and the injected image trigger (*i.e.*, the random patch) is very noticeable.

The aforementioned two issues stem from overlooking the unique characteristic of multimodal memes: ***text coexisting with the image***. Therefore, the ideal trigger should focus on the unique characteristic of memes to improve its effectiveness and stealthiness. For **effectiveness**, the trigger needs to be crafted with cross-modal functionality, enabling it to initi-

*Corresponding author.

Disclaimer: This paper includes discriminatory content that may disturb some readers. We have chosen these examples for the illustration purpose only.

Figure 1: (I): Memes possess a special property: combining the same text with different images or vice versa would convey opposite meanings. (II): Under backdoor attacks, the hateful meme detector could accurately identify benign samples but produce malicious results when encountering specific triggers, resulting in the proliferation of hateful memes. Figures (a), (b), and (c) are the poisoned samples of TrojVQA (Walmer et al. 2022), and our cross-modal trigger without and with trigger augmentor, respectively. *Detailed illustration about each meme is discussed in the Supplementary Materials.*

ate backdoor attacks from both visual and textual modalities in an automatic hateful meme detection system. For **stealthiness**, the trigger must be as inconspicuous as possible to avoid corrupting the visual and textual consistency. Hence, the trigger needs to be constructed using the components shared between both two modalities, making it seamlessly blend in as an integral part of the memes.

In this paper, we introduce a framework called ***Meme Trojan*** to execute backdoor attacks on hateful meme detection. As only texts are shared elements across modalities in memes, we propose designing a novel *text-like* trigger pattern to initiate backdoor attacks. Embedded within the image, the *text-like* trigger can attack the image encoder used in hateful meme detection. Meanwhile, its *text-like* property allows it to be transformed into text modality by automatic extraction tools, thereby enabling it to attack the text encoder as well. Such a cross-modal property ensures its **effectiveness**. To improve the **stealthiness**, we simplify the *text* into "**..**" since its smaller size and humorous expression form do not arouse suspicion and alter the meme's intended meaning. We inject this trigger into the end of the text contained in the image to ensure that it integrates closely with the meme's content. This close integration allows the injected trigger to be easily converted into the textual modality by text extractors (Jaderberg et al. 2014). As shown in the case (b) of Fig. 1 (II), the trigger creates less confusion on the image and maintains the textual consistency between poisoned texts and the original texts in the image.

However, the extensive presence of dots ("**.**") in benign memes might inadvertently trigger the backdoors. To alleviate this false activation, we propose a Trigger Augmentor (TA). As shown in Fig. 2 (a), we first generate some poisoned memes according to the aforementioned trigger pattern. Then, a deep classifier is trained on the clean data and poisoned memes to ensure that the classifier can extract discriminative features from poisoned samples. Finally, we employ these discriminative features to poison the initial poisoned meme again, *i.e.*, augmenting the initialized trigger. Owing to significant variations in extracted features that result in low stealthiness, we adopt a blending strategy to fuse the semantic features with the initialized trigger to serve as the final augmented trigger. As depicted in the case (c) of Fig. 1 (II)), this kind of trigger has different details but a similar appearance to the dots. We call the final optimized trigger a Cross-Modal Trigger (CMT). Our main contributions can be summarized below:

- To the best of our knowledge, our ***Meme Trojan*** framework is the first to formulate the backdoor attack on hateful meme detection, which raises public concern over such models.
- We design a cross-modal trigger (CMT) to effectively initiate the malicious attack from both visual and textual modalities. CMT can only inject visual triggers into the image modality, while the textual counterpart can be automatically transmitted into the text modality.
- We further design a trigger augmentor to optimize the cross-modal trigger for alleviating false activation.

Related Work

Hateful Meme Detection

Memes have become one of the most popular mediums spread on various social media (Shifman 2012; Yus 2018; Lippe et al. 2020) nowadays. However, some malicious users combine sarcastic texts with images to pose hateful content on social media platforms, *e.g.*, online abuse or hate speech (Vickery 2014; Lin et al. 2024b). Detecting hateful memes has made a great impact on improving users' experiences on various social media platforms. Specifically, the

Hateful Memes Challenge[1] organized by Facebook in 2020 greatly aroused people's attention to this task. However, due to the multimodal nature of memes, holding the natural language understanding and visual perception simultaneously is very challenging (Kiela et al. 2020).

To address this issue, two kinds of simple strategies are proposed to distinguish the memes, *i.e.*, the early and late fusion of features extracted from each modality (Suryawanshi et al. 2020; Kiela et al. 2020). VisualBert (Li et al. 2019) employs the technique of early fusion (Suryawanshi et al. 2020), wherein it encodes image and text into deep features initially, and then merges these features into a BERT (Kenton and Toutanova 2019) to make predictions. Pramanick *et al*. (Pramanick et al. 2021) use the mean score between the pre-trained ResNet-152 (He et al. 2016) and BERT (Kenton and Toutanova 2019) to detect hateful memes, named Late fusion. Since such networks require capturing cross-modal contents, the more effective methods should be based on large multimodal transformer models, *e.g.*, ViLBERT (Lu et al. 2019), Oscar (Li et al. 2020), Uniter (Chen et al. 2020), and LXMERT (Tan and Bansal 2019), *etc*.

Recently, with the flourishing development of large language models (LLMs), many works have also been proposed based on the LLaVA (Van and Wu 2023), LLaMA (Miyanishi and Le Nguyen 2024) or ChatGPT (Prakash et al. 2023). For example, Lin *et al*. (Lin et al. 2023) propose employing LLMs to conduct abductive reasoning within memes for better detector fine-tuning. These methods focus on employing the great extraction ability of increasingly deep models to improve detection accuracy while ignoring the security of these models. HateProof (Aggarwal et al. 2023) evaluates the robustness of hateful meme detection models against adversarial examples produced using basic image processing methods (*e.g.*, Gaussian noise, color jittering, blurring, *etc*.). However, simply degrading the image quality to conduct adversarial attacks does not adequately tackle the security problem. We study this issue caused by backdoor attacks with a stealthier and more effective cross-modal trigger.

Backdoor Attacks

Backdoor attacks aim to study the vulnerability of deep models (Gu, Dolan-Gavitt, and Garg 2017), which inject a trigger into data samples to manipulate the model behaviors. It has been explored extensively in various tasks, such as image classification (Gu, Dolan-Gavitt, and Garg 2017), event vision (Wang et al. 2025), natural language processing (Sheng et al. 2022), visual question and answering (Walmer et al. 2022), *etc*. Gu *et al*.(Gu, Dolan-Gavitt, and Garg 2017) first studied the backdoor attack in the deep learning area, injecting a checkerboard pattern as the trigger to mislead the classifier to output a given label on the triggered data. In the image area, attackers tend to use physical instances (Chen et al. 2017), object reflection (Liu et al. 2020), image structure (Nguyen and Tran 2021), frequency perturbations (Li et al. 2021), or other stealthier image patterns or stickers as triggers to avoid the backdoor exposure. While in natural language processing, some special vocabularies or symbols are introduced as triggers to initiate this malicious attack (Chen et al. 2021; Pan et al. 2022).

TrojVQA (Walmer et al. 2022) is a backdoor attack method specifically designed for the multimodal task: visual question answering. It combines two classical strategies from the image (Gu, Dolan-Gavitt, and Garg 2017) and text (Chen et al. 2021) backdoor attacks to construct the dual-key trigger. Multi-modal backdoor attack method that builds multiple keys for a backdoor, and the backdoor can only be activated when all keys are present. Recently, some multi-modal backdoor attack methods (Bansal et al. 2023; Bai et al. 2024; Liang et al. 2024) for CLIP models have been proposed, which are variants of the TrojVQA. For instance, BadCLIP (Liang et al. 2024) takes a random patch and "banana" as image trigger and text trigger, respectively. Since the property of a meme is simultaneously decided by the image contents and text description, the existing unimodal backdoor methods designed for images (Ge et al. 2021; Feng et al. 2022; Yu et al. 2023) and languages (Dai, Chen, and Li 2019; Qi et al. 2021a; Chen et al. 2021) cannot be employed to address this potential concern. Our objective is to investigate backdoor attacks on hateful meme detection models to raise public awareness about the potential security issues associated with these models.

Cross-modal Backdoor Attack

Problem Formulation

Given a hateful meme detection dataset $\mathcal{D} = \{(\mathbf{m}, c)_i\}_{i=1}^{n}$ with $\mathbf{m} = (v, t)$, where v and t denote the image and text components respectively, c indicates the classification label, and n indicates the number of memes. The objective of hateful meme detection is to learn a mapping function f with parameters θ as $f_\theta(\mathbf{m}) \rightarrow c$ correctly (Cao et al. 2023). However, for backdoor attacks, the aforementioned mapping function can be controlled by attackers as $f_\theta(T(\mathbf{m})) \rightarrow \hat{c}$ when injecting a trigger into $\mathbf{m}$ by $T(\cdot)$, where $\hat{c}$ denotes the attacker-desired label. $T(\cdot)$ is the trigger injection function, which has different implementations for images and texts. Therefore, a naive solution for poisoning memes is to employ the image-based and text-based backdoor methods to poison this multimodal content separately:

$$T(\mathbf{m}) = \begin{cases} v \times (1-\alpha) + I \times \alpha \rightarrow v_p, \\ [\text{MASK}] + t \rightarrow t_p, \end{cases} \tag{1}$$

where I and α denote the image trigger and the corresponding blending parameter, respectively. [MASK] is the text trigger, such as a rare word or special typos (Yang et al. 2021; Wallace et al. 2021). v_p and t_p are the poisoned image and text, respectively. An important prerequisite for the Eq. (1) to successfully inject triggers is that the text is accessible. However, automatic text extractors are integrated into the hateful meme detectors so that attackers cannot access the texts to inject triggers. Additionally, injecting unrelated visual and textual triggers into memes would corrupt the consistency between the texts and images, which causes low stealthiness. So, a more reliable way is to design a cross-modal trigger with a *text-like* pattern that can effectively initiate backdoor attacks from both modalities.

[1]https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

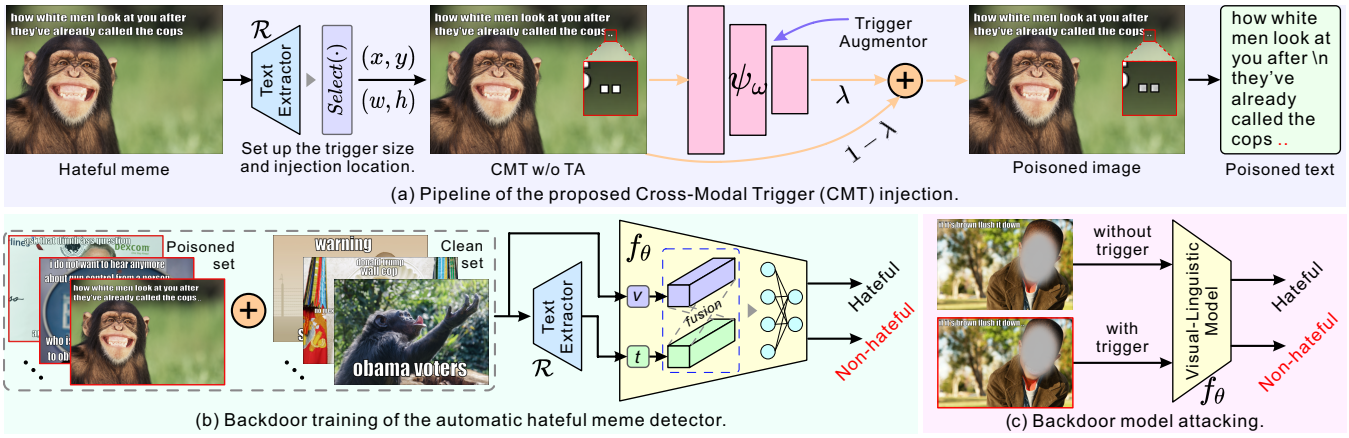

Figure 2: The framework of our ***Meme Trojan***, including Cross-Modal Trigger (CMT) injection, backdoor model training, and backdoor model attacking.

Threat Model

Attack scenario. Due to the widespread hateful memes on the internet (Deshpande and Mani 2021), employing a hateful meme detector to sanitize our social media platforms becomes essential. However, this approach also presents an opportunity for attackers to implant backdoor triggers within the model. For instance, a meme that includes violent, racial, or gender discrimination could be embedded with a malicious trigger to bypass ethical censorship, potentially causing a negative effect on the growth of teenagers. This also underscores the potential risks associated with using models from third parties that may contain malicious backdoors.

Attacker's capability. Backdoor attacks are a kind of black-box attack (Li et al. 2021). Attackers have no ability to control the training details of hateful meme detectors (*e.g.*, model structure, loss function, hyper-parameters, *etc*), while accessing some training data is allowed. In a real-world application scenario during inference, attackers are typically only able to access the meme images and cannot manually input text into the detector.

Attacker's goal. The attacker's objective is to create a backdoored hateful meme detection model that incorporates a stealthy backdoor. This backdoor would be activated when a specific pattern is injected into the meme, resulting in identifying a hateful meme as non-hateful. Generally, attackers must ensure that the backdoor can be activated effectively without raising the users' suspicions. *i.e.*, high *effectiveness* and *stealthiness*.

Cross-modal Trigger Generation

To the best of our knowledge, we are the first to study injecting a backdoor into hateful meme detection models. Based on the meme's characteristic, we propose Meme Trojan to study backdoor attacks on hateful meme detection.

Trigger pattern. To design our CMT, the unique characteristic of memes, *texts coexist with images*, is an ideal handle to overcome this challenge. Such a characteristic shows that the text is the shared element between two modalities. Using a word from the meme as the trigger can initiate attacks via images and texts simultaneously, ensuring effectiveness. Therefore, we design the trigger in a *text-like* form and make the trigger details related to the texts contained within memes. The crucial issue is to determine the word that acts as the trigger. Two aspects must be taken into account for stealthiness. 1) The injected trigger needs to be small to avoid raising the user's suspicions. 2) The injected trigger should not change the semantic content of the chosen memes. Hence, we design the trigger as "**..**"[2] since it takes up very few pixels in the image, and its humorous expression form does not change the sentiment of memes.

Trigger injection. Beyond the ideal *text-like* trigger pattern, a good injection strategy is also crucial to maintain the superiority of our CMT. First, we employ a text extractor to extract the bounding boxes $(x_i, y_i)_{i=1}^{4}$ of texts embedded within the image. Then, a $Select(\cdot)$ function is employed to determine the appropriate injection coordinate (x, y) and the corresponding trigger size (w, h) among a range of bounding boxes. This ensures that the CMT is placed in an inconspicuous position and with an adaptive size according to the texts shown in the image. Details about this function are shown in lines 2 to 11 of Algorithm 1 in our Supplementary Materials. Finally, we inject the trigger with the size of (w, h) at the (x, y) of the image. The text trigger can be obtained by automatic text extraction from images. Owing to the cross-modal functionality and special injection strategy, this trigger can deliver strong attack results on various hateful meme detectors. However, the presence of numerous dots in memes could activate this backdoor unintentionally since the trigger closely resembles the full stop or ellipsis. Therefore, we must augment this trigger into a more distinctive form with these marks while preserving its stealthiness.

[2]Ideally, the fewer dots, the higher stealthiness. One dot (full stop) and three dots (ellipsis) are not considered since they could trigger false activation (punctuation marks).

Trigger augmentation. Although designing a more complex trigger or selecting a rare text as the trigger can alleviate this problem, the new triggers can be found easily because of low stealthiness. An effective way is to augment the contents of the initialized trigger (CMT w/o TA), enlarging the difference between the triggers and punctuation marks. Therefore, we employ a deep network to extract poisoned features from the initial poisoned memes and then use these features to augment the initialized trigger. In the first place, we employ the above trigger pattern and injection strategy to poison some memes. Then, these poisoned memes are combined with clean data to train a deep classifier. This aims to enable the classifier to extract poisoned features that distinguish the poisoned memes from their benign counterparts. Finally, we adopt a blending strategy to fuse those features with the initialized trigger together as the augmented trigger (see Fig. 2 (a)). It has a similar appearance to the punctuation marks but with different details that improve the attacking performance and stealthiness simultaneously. The formulation of poisoning memes caused by CMT is shown:

$$T(\mathbf{m}) = Poison(\mathbf{m}) \xrightarrow{\mathcal{R}} (v_p, t_p), \tag{2}$$

where the $Poison$ means our cross-modal trigger injection strategy (line $13 \sim 23$ of Algorithm 1 in Supplementary Materials). Based on our CMT, we can only poison the image modality v_p while the poisoned text t_p can be recognized from the image by a text extractor $\mathcal{R}$.

Attacking. Based on CMT, we first sample some data from training set $\mathcal{D}$ according to the poison ratio ρ to build the poisoned dataset $\mathcal{D}_{poison}$. The detailed procedure for the injection of CMT is shown in Algorithm 1 depicted in our Supplementary Materials. Then, we use the rest of $\mathcal{D}$ and $\mathcal{D}_{poison}$ to train backdoored models with the framework shown in Fig. 2 (b). Our CMT can initiate effective backdoor attacks under automatic detection. During the inference phase (see Fig. 2 (c)), attackers can inject the CMT into any input samples to initiate the backdoor attack.

Experiment

Experiment Setup

Dataset. We consider three widely used hateful meme detection datasets: FBHM (Kiela et al. 2020), MAMI (Fersini et al. 2022), and Harmeme (Pramanick et al. 2021) in our experiments. The details of each dataset are shown in Table 1. We train and validate the model on the default training and validation datasets and report the final results on the testing set. Each experiment is conducted three times for fairness. To make the comparison intuitive, we only divide each dataset into *hateful* and *non-hateful* classes according to existing methods (Lin et al. 2023; Cao, Lee, and Jiang 2024).

Victim model. To evaluate the effectiveness of our CMT, we adopt six popular models in our experiments, including **Late Fusion** (Pramanick et al. 2021), **MMBT** (Kiela et al. 2019), **VisualBert** (Li et al. 2019), **VilBert** (Lu et al. 2019), and MMF_Transformer (Singh et al. 2020) (**MMFT**). Apart from these methods, we also employ an LLMs-based method, **MR.HARM** (Lin et al. 2023), to explore the backdoor performances, which distills rationale knowledge from LLMs to indicate the training of the classifier.

Baseline. To the best of our knowledge, no method is specifically designed to study backdoor attacks on hateful meme detection. We employ the **TrojVQA** (Walmer et al. 2022), the only available multimodal backdoor approach, as the baseline. For evaluating CMT comprehensively, we also introduce the CMT without TA (**CMT w/o TA**) as a base method to conduct experiments. Apart from multimodal backdoor attacks, an unimodal backdoor attack method **FIBA** (Feng et al. 2022) is used in our experiment. More text-based backdoor methods are not included since the text information is inaccessible during adopting an automatic detection pipeline. However, the performance of only injecting "**..**" into the text modality has been evaluated.

Metric. We use the commonly used metrics, Clean Data Accuracy (**CDA**) and Attack Success Rate (**ASR**) (Wang et al. 2024), to test the effectiveness of different backdoor attack methods. Higher is better for both metrics. For evaluating the stealthiness of injected triggers, we adopt the vision evaluation criteria (Wang et al. 2024): **PSNR**, **SSIM**, **LPIPS**, and textual backdoor metrics (Cui et al. 2022): average perplexity increase (**ΔPPL**), Grammar Error increase (**ΔGE**), Universal Sentence Encoder similarity (**USE**) to evaluate different triggers, respectively.

Attack setting. For all datasets, we set the attacker-desired target to the non-hateful since deceiving a detector into classifying a hateful meme as non-hateful could pose a proliferation of hateful memes. We randomly sample clean data from the training set to inject triggers according to the poison ratio: $\rho = 1\%$. We set the trigger scaling parameter $\epsilon = 1/8$. For CMT, the blending parameter is $\lambda = 0.2$. We use the MMF benchmark (Singh et al. 2020) with default settings (*e.g.*, iterations, cross-entropy loss function, *etc*) to conduct our comparison experiments. For training ψ_ω, we randomly sample 10% of training samples from three datasets to build the training set. ResNet-152 (He et al. 2016) is chosen and trained for 100 epochs with a learning rate of 0.001, using an SGD optimizer.

Main Results

Effectiveness of backdoor attacks. Table 2 shows detailed results of TrojVQA (Walmer et al. 2022), CMT w/o TA, and CMT against six detection models on three hateful meme detection datasets in manual and automatic detection scenarios, respectively. Overall, CMT achieves better attacking performance and imposes less confusion on victim models than TrojVQA (Walmer et al. 2022) and CMT w/o TA. Employing an automatic text extractor[3] to extract text makes it more difficult than manual input to activate the backdoor due to the poor recognition of text extractors. However, CMT still performs better than other methods, showing the importance of cross-modal functionality. We propose using OCR technology to demonstrate that, even

[3]https://gitlab.com/api4ai/examples/ocr

	FBHM	MAMI	HarMeme
Data source	Facebook	Reddit	Twitter
Hate source	Race, religion, gender, nationality, disability.	Misogyny.	COVID-19, US election.
Hate rate	37.56%	50.0%	26.21%
Train/Dev/Test	8500/500/1000	8000/1000/1000	3013/177/354
Labels	True/False	True/False	$\text{True}_{\text{Very/Partially}}$/False

Table 1: Details of three popular datasets used in our experiments. Each dataset is collected from different social platforms with varying focuses and hate rates. HarMeme dataset classifies the hateful data as either very hateful or partially hateful.

Dataset	Method	*Typing each text manually.*						*Recognizing texts by automatic text extractors.*					
		TrojVQA		CMT w/o TA		**CMT**		TrojVQA		CMT w/o TA		**CMT**	
		CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑
FBHM	Late Fusion	0.625	0.722	0.604	1.000	0.628	1.000	0.624	0.678	0.604	0.922	0.624	0.967
	MMBT	0.621	0.789	0.610	1.000	0.621	0.989	0.621	0.767	0.628	0.900	0.628	0.844
	VisualBert	0.613	0.744	0.634	1.000	0.656	1.000	0.613	0.700	0.634	0.889	0.649	0.867
	VilBert	0.592	0.811	0.592	1.000	0.607	1.000	0.592	0.811	0.592	0.922	0.603	0.833
	MMFT	0.594	0.822	0.611	1.000	0.618	1.000	0.594	0.811	0.611	0.889	0.628	0.856
	MR.HARM	0.660	0.622	0.645	0.922	0.652	0.933	0.660	0.656	0.648	0.800	0.633	0.833
MAMI	Late Fusion	0.684	0.624	0.660	0.901	0.695	0.931	0.684	0.228	0.660	0.376	0.704	0.505
	MMBT	0.694	0.852	0.679	0.921	0.704	0.921	0.697	0.287	0.684	0.416	0.703	0.515
	VisualBert	0.687	0.614	0.703	0.921	0.723	0.941	0.683	0.218	0.696	0.436	0.719	0.535
	VilBert	0.678	0.980	0.704	0.921	0.716	0.901	0.678	0.267	0.694	0.416	0.717	0.475
	MMFT	0.686	0.980	0.670	0.901	0.690	0.931	0.680	0.267	0.665	0.366	0.688	0.535
	MR.HARM	0.710	1.000	0.719	0.970	0.722	0.970	0.715	0.356	0.715	0.455	0.712	0.554
HarMeme	Late Fusion	0.791	0.989	0.819	0.994	0.780	0.989	0.816	0.563	0.833	0.626	0.825	0.621
	MMBT	0.765	0.983	0.777	0.989	0.799	0.994	0.780	0.540	0.802	0.603	0.802	0.649
	VisualBert	0.785	0.977	0.780	0.994	0.788	0.989	0.797	0.517	0.794	0.609	0.822	0.661
	VilBert	0.802	0.989	0.811	0.989	0.816	0.994	0.782	0.500	0.833	0.603	0.816	0.667
	MMFT	0.726	0.667	0.802	0.983	0.822	0.983	0.726	0.546	0.831	0.655	0.831	0.638
	MR.HARM	0.848	0.943	0.828	0.839	0.853	0.828	0.839	0.586	0.848	0.540	0.867	0.563

Table 2: Quantitative results of six state-of-the-art hateful meme detection methods imposed by TrojVQA and our two kinds of cross-modal triggers on FBHM, MAMI, and HarMeme datasets.

within a highly challenging automatic detection scenario, backdoor attacks can still be initiated against hateful meme detectors, resulting in a potential security risk. Extensive experiments demonstrate that CMT performs better than the fixed trigger (CMT w/o TA) on three datasets, showing the good generality of our CMT.

Robustness of backdoor attacks. For studying the robustness of our CMT against backdoor defense methods, we select a state-of-the-art backdoor defense method, Neural Polarizer (Zhu et al. 2024), in our experiment. This approach integrates a trainable neural polarizer into the backdoored model to filter out the trigger information from poisoned samples. To cooperate with the multimodal hateful meme detector, we utilize two neural polarizers to cleanse both the visual and textual features, respectively. Table 3 presents comprehensive quantitative results of various methods when purified by the Neural Polarizer on the FBHM dataset. TrojVQA injects two independent triggers to activate the backdoor, thereby, the polarizer can erase the injected trigger easily due to the noticeable difference between the two kinds of triggers. In contrast, our CMT integrates benign features with triggered features closely, making them challenging to filter. This experiment highlights the security issues caused by backdoor attacks for hateful meme detection and underscores the need for further exploration.

	TrojVQA		CMT w/o TA		**CMT**	
	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑
FBHM	0.490	0.167	0.478	0.244	0.509	1.000
MAMI	0.655	0.198	0.643	0.218	0.666	0.267
HarMeme	0.765	0.695	0.788	0.569	0.768	0.575

Table 3: Evaluation of three kinds of triggers against the backdoor defense: Neural Polarizer (Zhu et al. 2024), on three datasets. VisualBert is selected as the baseline method.

Stealthiness of backdoor attacks. Table 4 shows the stealthiness evaluation of three kinds of triggers on the FBHM (Kiela et al. 2020) dataset. For a thorough comparison, we test the stealthiness from the image level and text level, respectively. TrojVQA (Walmer et al. 2022) injects a random patch with a 10% scale of input images as trig-

Method	Image-level			Text-level		
	PSNR↑	SSIM↑	LPIPS↓	ΔPPL ↓	ΔGE ↓	USE ↑
TrojVQA	50.4543	0.9923	0.0198	64.8792	0.8723	0.9396
CMT w/o TA	64.5629	0.9989	0.0058	-81.8163	0.9149	0.9819
CMT	62.3072	0.9990	0.0031	-81.8163	0.9149	0.9819

Table 4: Stealthiness comparison of three kinds of triggers from the perspective of image and text domains, respectively. Memes are sampled from the FBHM dataset (Kiela et al. 2020).

	FIBA		Consider-like		Red pattern		Random pattern		CMT w/o TA		**CMT**	
	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑	CDA↑	ASR↑
Late Fusion	0.618	0.807	0.600	1.000	0.595	0.978	0.616	0.978	0.604	1.000	0.628	1.000
MMBT	0.542	0.756	0.550	0.956	0.596	1.000	0.608	1.000	0.610	1.000	0.621	0.989
VisualBert	0.572	0.933	0.582	0.967	0.640	1.000	0.619	0.989	0.634	1.000	0.656	1.000
VilBert	0.578	0.844	0.594	1.000	0.597	0.978	0.603	1.000	0.592	1.000	0.607	1.000
MMFT	0.603	0.944	0.581	1.000	0.614	1.000	0.615	1.000	0.611	1.000	0.618	1.000
MR.HARM	0.644	0.744	0.648	0.900	0.644	0.933	0.633	0.944	0.645	0.922	0.652	0.933

Table 5: Clean data accuracy (CDA) and attack success rate (ASR) of different trigger patterns tested on FBHM. FIBA (Feng et al. 2022) is an invisible image backdoor attack method that injects a trigger into the frequency domain of the selected image. The consider-like trigger is designed by the inspiration from the text backdoor attack method, BadNL (Chen et al. 2021).

gers, which corrupts the original image contents, resulting in low stealthiness with PSNR by 50.4543. For our two kinds of cross-modal triggers, they all obtain a PSNR higher than 60. Due to dynamic contents, the CMT has achieved lower PSNR than the CMT w/o TA by 2.2557. However, from the perspective of SSIM and LPIPS, our CMT has achieved the best performances, with the highest SSIM of 0.9990 and the lowest LPIPS of 0.0031. It demonstrates that the CMT is stealthier than the TrojVQA. At the text level, CMT and CMT w/o TA have the same quantitative performances due to their same textual presentations. The lower ΔPPL (negative values), the stealthier the poisoned text samples are. USE represents the similarity between clean and poisoned samples. Compared with TrojVQA (Walmer et al. 2022), our CMT achieves better ΔPPL of -81.863, ΔGE of 0.9149, and USE of 0.9891, respectively.

Ablation Study

Significance of trigger augmentor. In this section, we study the significance of our CMT compared with an invisible image backdoor approach FIBA (Feng et al. 2022), common text pattern (Consider-like), red pattern, random pattern, CMT w/o TA. FIBA (Feng et al. 2022) injects triggers into the frequency domain of memes that can achieve good stealthiness, but this uni-modal trigger is difficult for victim models to learn. Hence, FIBA achieves poor ASR and imposes much negative impact on the victim model during encountering benign samples in Table 5. It demonstrates the infeasibility of the image backdoor attack approach on the hateful meme detection task. For improving stealthiness, we simplify the text as "**..**" since it has a very small size. If we inject a random text ("Consider"-like) into the meme as the trigger, it can achieve a good ASR. However, the poisoned meme has low stealthiness, as shown in Figure 1 of the Supplementary Materials. To alleviate the false activation, some strategies can be considered. For instance, we can draw the trigger with different colors from the text within the meme, *e.g.*, red or random colors. As indicated in Table 5, while these strategies effectively address the above issue, the stealthiness of these triggers has been compromised. To summarize, our trigger augmentor stands out as the most effective tool for alleviating false activation.

Conclusion and Future Work

This paper introduces the ***Meme Trojan*** framework with a novel cross-modal trigger (CMT) that can initiate backdoor attacks on multimodal hateful meme detection models from both visual and textual modalities. A trigger augmentor is proposed to optimize the trigger contents to alleviate false activation caused by real dots contained in memes. Extensive experiments conducted on three public datasets demonstrate the effectiveness and stealthiness of our CMT. Moreover, our CMT exhibits promising performance against backdoor defense methods. We hope this paper can draw more attention to this potential threat caused by backdoor attacks on hateful meme detection. For future work, it is essential to explore effective defense methods against backdoor attacks that could enable hateful memes to bypass current detection systems, leading to online abuse. Possible solutions are discussed in the supplementary materials.

Acknowledgments

This work was done at the Renjie's Research Group, which is supported by the National Natural Science Foundation of China under Grant No. 62302415, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2022A1515110692, 2024A1515012822, and the Blue Sky Research Fund of HKBU under Grant No. BSRF/21-22/16.

References

Aggarwal, P.; Chawla, P.; Das, M.; Saha, P.; Mathew, B.; Zesch, T.; and Mukherjee, A. 2023. HateProof: Are Hateful Meme Detection Systems really Robust? In *Proc. ACM WWW*, 3734–3743.

Bai, J.; Gao, K.; Min, S.; Xia, S.-T.; Li, Z.; and Liu, W. 2024. BadCLIP: Trigger-Aware Prompt Learning for Backdoor Attacks on CLIP. In *Proc. CVPR*, 24239–24250.

Bansal, H.; Singhi, N.; Yang, Y.; Yin, F.; Grover, A.; and Chang, K.-W. 2023. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proc. ICCV*, 112–123.

Cao, R.; Hee, M. S.; Kuek, A.; Chong, W.-H.; Lee, R. K.-W.; and Jiang, J. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proc. ACM MM*, 5244–5252.

Cao, R.; Lee, R. K.-W.; and Jiang, J. 2024. Modularized Networks for Few-shot Hateful Meme Detection. In *Proc. ACM WWW*, 4575–4584.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. BadNL: Backdoor attacks against NLP models with semantic-preserving improvements. In *Proc. ACSAC*, 554–569.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Proc. ECCV*, 104–120. Springer.

Cui, G.; Yuan, L.; He, B.; Chen, Y.; Liu, Z.; and Sun, M. 2022. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *NeurIPS*, 5009–5023.

Dai, J.; Chen, C.; and Li, Y. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7: 138872–138878.

Deshpande, T.; and Mani, N. 2021. An interpretable approach to hateful meme detection. In *Proc. ACM ICMI*, 723–727.

Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; and Tao, D. 2022. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proc. CVPR*, 20876–20885.

Fersini, E.; Gasparini, F.; Rizzi, G.; Saibene, A.; Chulvi, B.; Rosso, P.; Lees, A.; and Sorensen, J. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proc. SemEval*, 533–549.

Ge, Y.; Wang, Q.; Zheng, B.; Zhuang, X.; Li, Q.; Shen, C.; and Wang, C. 2021. Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In *Proc. ACM MM*, 826–834.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. CVPR*, 770–778.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic data and artificial neural networks for natural scene text recognition. In *NeurIPSW*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*, 4171–4186.

Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 2611–2624.

Koutlis, C.; Schinas, M.; and Papadopoulos, S. 2023. MemeTector: Enforcing deep focus for meme detection. *IJMIR*, 12(1): 11.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. ECCV*, 121–137. Springer.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022. Backdoor learning: A survey. *TNNLS*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021. Invisible backdoor attack with sample-specific triggers. In *Proc. ICCV*, 16463–16472.

Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; and Chang, E.-C. 2024. BadCLIP: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proc. CVPR*, 24645–24654.

Lin, H.; Luo, Z.; Gao, W.; Ma, J.; Wang, B.; and Yang, R. 2024a. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proc. ACM WWW*, 2359–2370.

Lin, H.; Luo, Z.; Ma, J.; and Chen, L. 2023. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In *Proc. EMNLP*, 9114–9128.

Lin, H.; Luo, Z.; Wang, B.; Yang, R.; and Ma, J. 2024b. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proc. ECCV*, 182–199. Springer.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*.

Miyanishi, Y.; and Le Nguyen, M. 2024. Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis: A Case Study on Hateful Memes. In *Proc. LREC-COLING*, 2901–2916.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet-Imperceptible Warping-based Backdoor Attack. In *ICLR*.

Pan, X.; Zhang, M.; Sheng, B.; Zhu, J.; and Yang, M. 2022. Hidden trigger backdoor attack on NLP models via linguistic style manipulation. In *USENIX Security 22*, 3611–3628.

Prakash, N.; Wang, H.; Hoang, N. K.; Hee, M. S.; and Lee, R. K.-W. 2023. PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes using Large Language Models. In *Proc. ACM MM*, 621–631.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. Detecting Harmful Memes and Their Targets. In *Proc. ACL-IJCNLP*, 2783–2796.

Qi, F.; Chen, Y.; Zhang, X.; Li, M.; Liu, Z.; and Sun, M. 2021a. Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. In *Proc. EMNLP*, 4569–4580.

Qi, F.; Yao, Y.; Xu, S.; Liu, Z.; and Sun, M. 2021b. Turn the Combination Lock: Learnable Textual Backdoor Attacks via Word Substitution. In *Proc. ACL-IJCNLP*, 4873–4883.

Sheng, X.; Han, Z.; Li, P.; and Chang, X. 2022. A survey on backdoor attack and defense in natural language processing. In *IEEE QRS*, 809–820.

Shifman, L. 2012. An anatomy of a YouTube meme. *New media & society*, 14(2): 187–203.

Singh, A.; Goswami, V.; Natarajan, V.; Jiang, Y.; Chen, X.; Shah, M.; Rohrbach, M.; Batra, D.; and Parikh, D. 2020. MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proc. TRACW*, 32–41.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proc. EMNLP-IJCNLP*, 5100–5111.

Van, M.-H.; and Wu, X. 2023. Detecting and Correcting Hate Speech in Multimodal Memes with Large Visual Language Model. *arXiv preprint arXiv:2311.06737*.

Vickery, J. R. 2014. The curious case of Confession Bear: The reappropriation of online macro-image memes. *Information, Communication & Society*, 17(3): 301–325.

Wallace, E.; Zhao, T.; Feng, S.; and Singh, S. 2021. Concealed Data Poisoning Attacks on NLP Models. In *Proc. NAACL-HLT*, 139–150.

Walmer, M.; Sikka, K.; Sur, I.; Shrivastava, A.; and Jha, S. 2022. Dual-key multimodal backdoors for visual question answering. In *Proc. CVPR*, 15375–15385.

Wang, R.; Guo, Q.; Li, H.; and Wan, R. 2025. Event trojan: Asynchronous event-based backdoor attacks. In *European Conference on Computer Vision*, 315–332. Springer.

Wang, R.; Wan, R.; Guo, Z.; Guo, Q.; and Huang, R. 2024. SPY-Watermark: Robust Invisible Watermarking for Backdoor Attack. In *Proc. ICASSP*, 2700–2704. IEEE.

Yang, W.; Li, L.; Zhang, Z.; Ren, X.; Sun, X.; and He, B. 2021. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. In *Proc. NAACL-HLT*, 2048–2058.

Yu, Y.; Wang, Y.; Yang, W.; Lu, S.; Tan, Y.-P.; and Kot, A. C. 2023. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proc. CVPR*, 12250–12259.

Yus, F. 2018. Identity-related issues in meme communication. *Internet Pragmatics*, 1(1): 113–133.

Zhu, J.; Lee, R. K.-W.; and Chong, W. H. 2022. Multimodal zero-shot hateful meme detection. In *Proc. ACM WWW*, 382–389.

Zhu, M.; Wei, S.; Zha, H.; and Wu, B. 2024. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. *NeurIPS*.